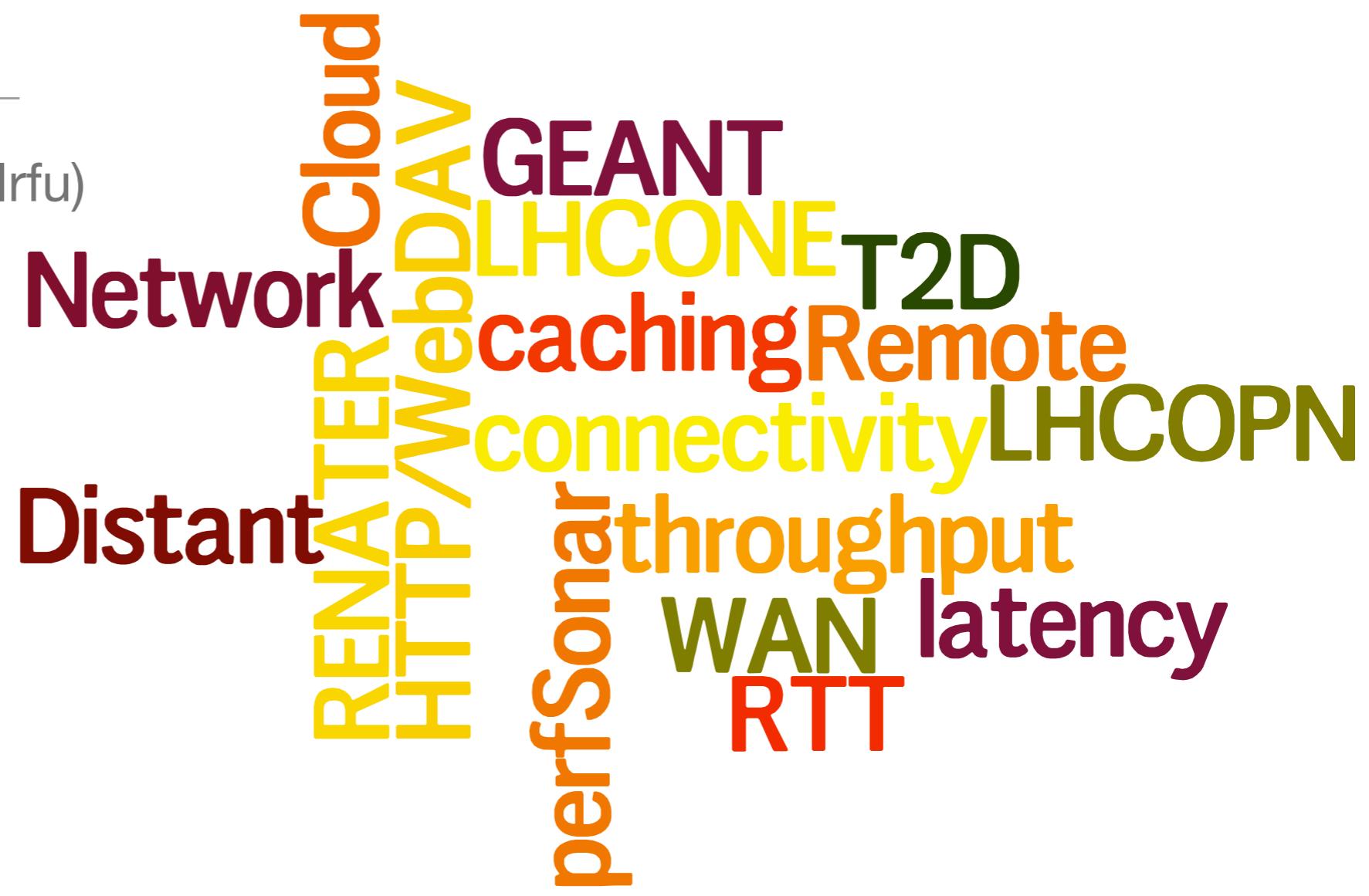


# Network issues on FR cloud

---

Eric Lançon (CEA-Saclay/Irfu)





## Network Usage

- Data distribution
- MC production
- Analysis
- Distributed storage

©2010 Google

# Network used for

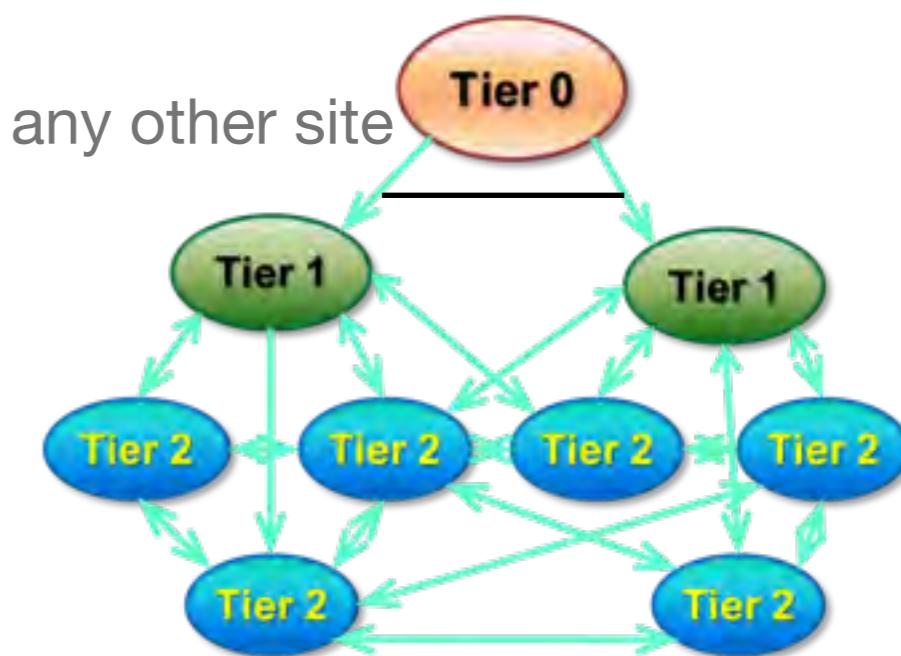
---

- **Data distribution**, 2 components :
  - Pre-placed data (a la MONARC)
  - Dynamic data distribution  
(popular data to available sites)
- **Analysis**
  - Retrieving results
  - Small data sets
- **Distributed storage**
  - To optimize resources
  - Simplify data management
- **MC production**
  - Within a given cloud
  - Across clouds

**The ‘old’ computing model is dying**

## The ATLAS Data Model has changed

- Moved away from the historical model
- 4 recurring themes:
  - **Flat(ter) hierarchy:** Any site can replicate data from any other site
  - **Multi Cloud Production**
    - Need to replicate output files to remote Tier-1
  - **Dynamic data caching:** Analysis sites receive datasets from any other site “on demand” based on usage pattern
    - Possibly in combination with pre-placement of data sets by centrally managed replication of datasets
  - **Remote data access:** local jobs accessing data stored at remote sites
- **ATLAS is now heavily relying on multi-domain networks and needs decent e2e network monitoring**



# ATLAS sites and connectivity

- ATLAS computing model has (will continue to) changed
  - More experience
  - More tools and monitoring
- New category of sites : Direct T2s (**T2Ds**)
  - Primary hosts for datasets (**analysis**) and for group analysis
  - Get and send data from different clouds
  - Participate in cross cloud production

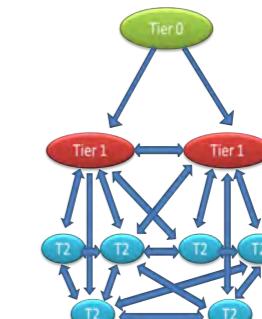
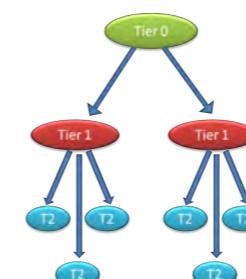
## T2D: revising the criteria

### New criteria - under evaluation

- All transfers from the candidate T2D to **9/12** T1s for big files ('L') must be above **5 MB/s** during the last week and during 3 out of the **5** last weeks.
- All transfers from **9/12** T1s to the candidate T2D for big files must be above **5 MB/s** during the last week and during 3 out of the **5** last weeks

<http://gnegri.web.cern.ch/gnegri/T2D/t2dStats.html>

**FR-cloud T2Ds** : BEIJING, GRIF-LAL, GRIF-LPNHE, IN2P3-CPPM, IN2P3-LAPP, IN2P3-LPC, IN2P3-LPSC



# Network performance monitoring

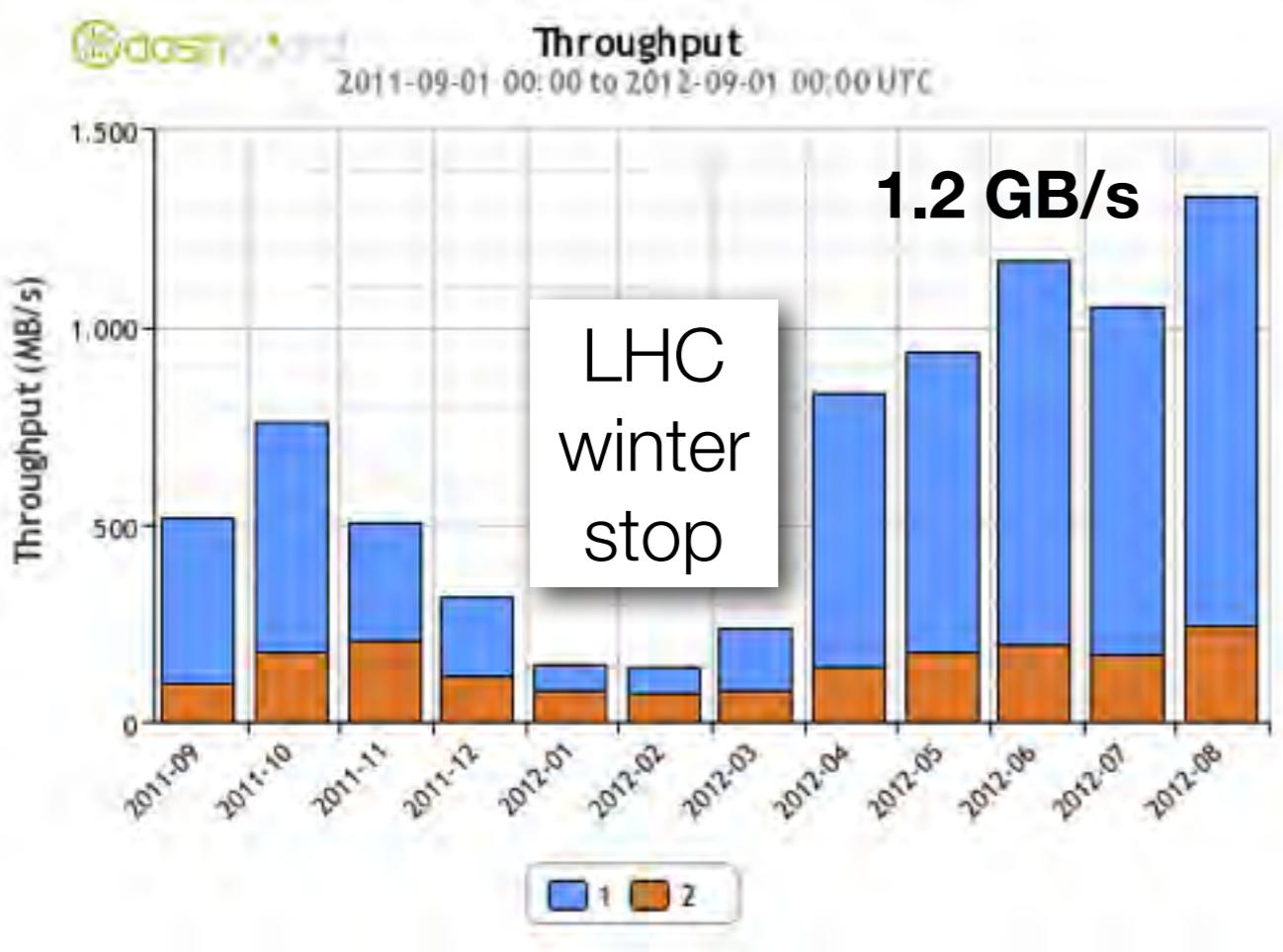
---

- **Networking accounting :**
  - **Organized** (FTS) file transfers : <http://dashb-atlas-data.cern.ch/ddm2/>, not for direct transfers by users (dq2-get)
- **ATLAS ‘sonar’ :**
  - Calibrated file transfers by ATLAS Data Distribution system, from **storage to storage** : <http://bourricot.cern.ch/dq2/ftsmon/>
  - > 1 GB file transfers used to monitor and validate T2Ds
- **perfSONAR (PS) :**
  - **Network performance** (throughput, latency) : <http://perfsonar.racf.bnl.gov:8080/exda/>
  - Located as close as possible to storage at site and with similar connection hardware

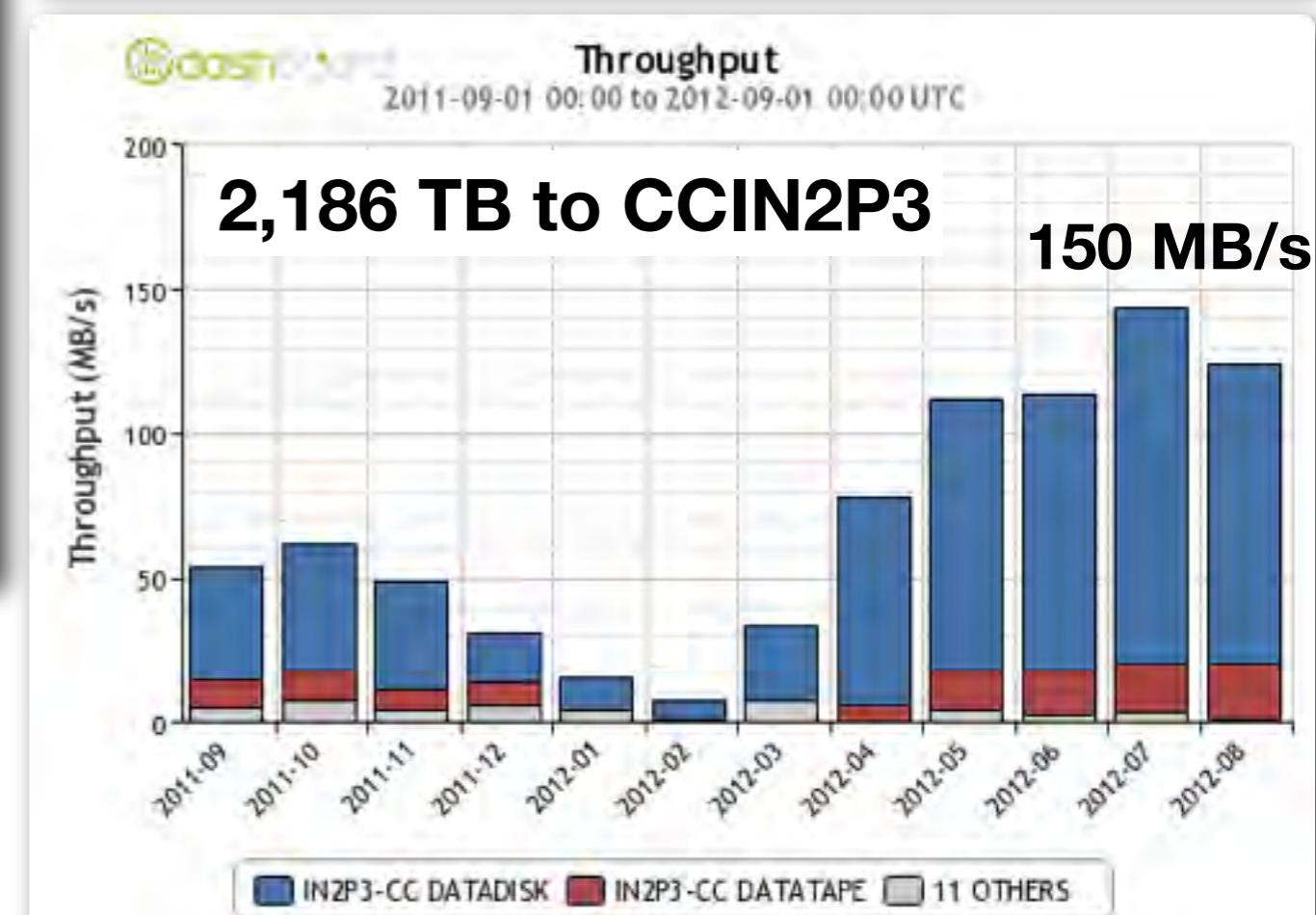
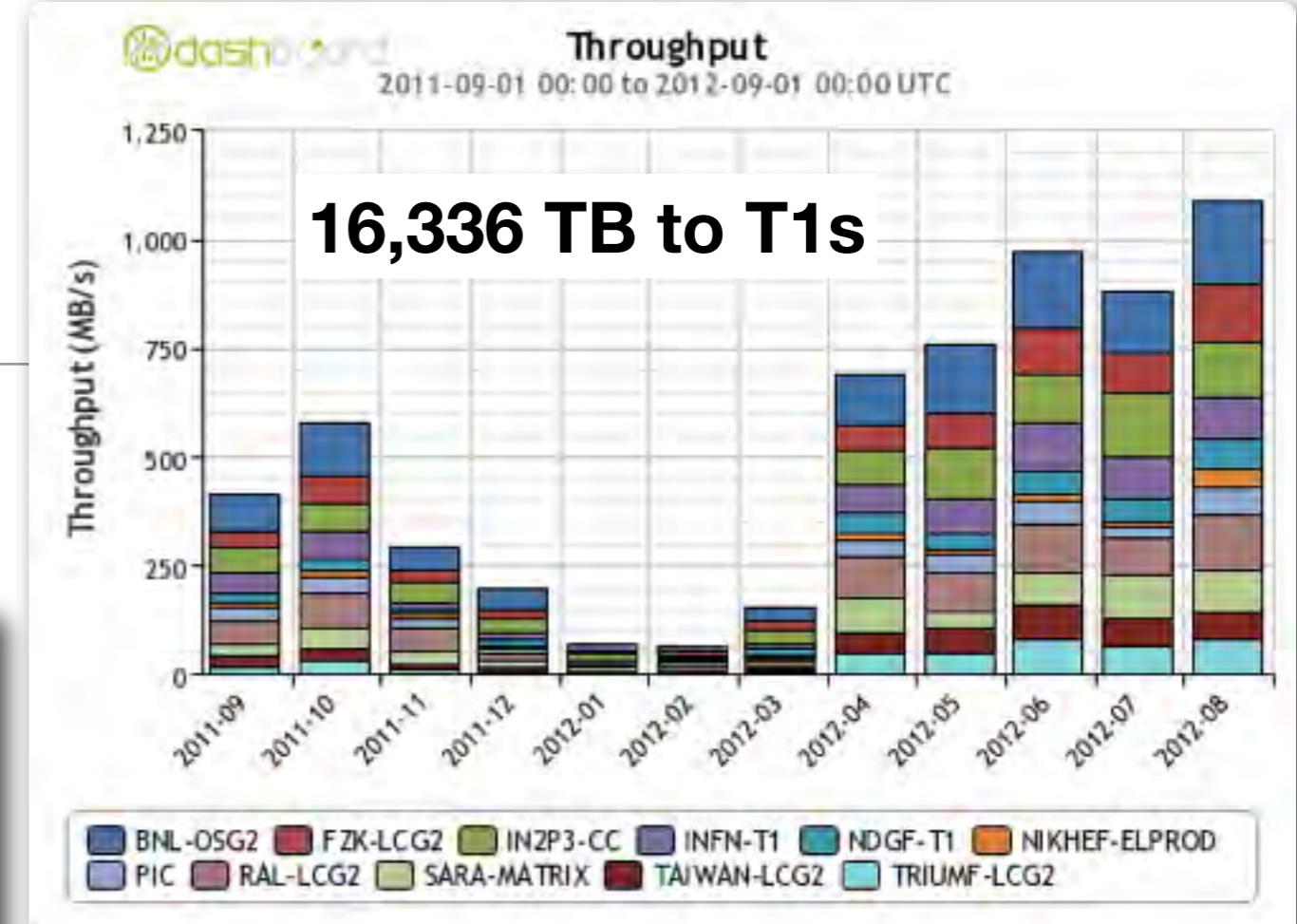
# T0 exports over a year

**Over 1GB/s**

Better LHC efficiency and higher trigger rate

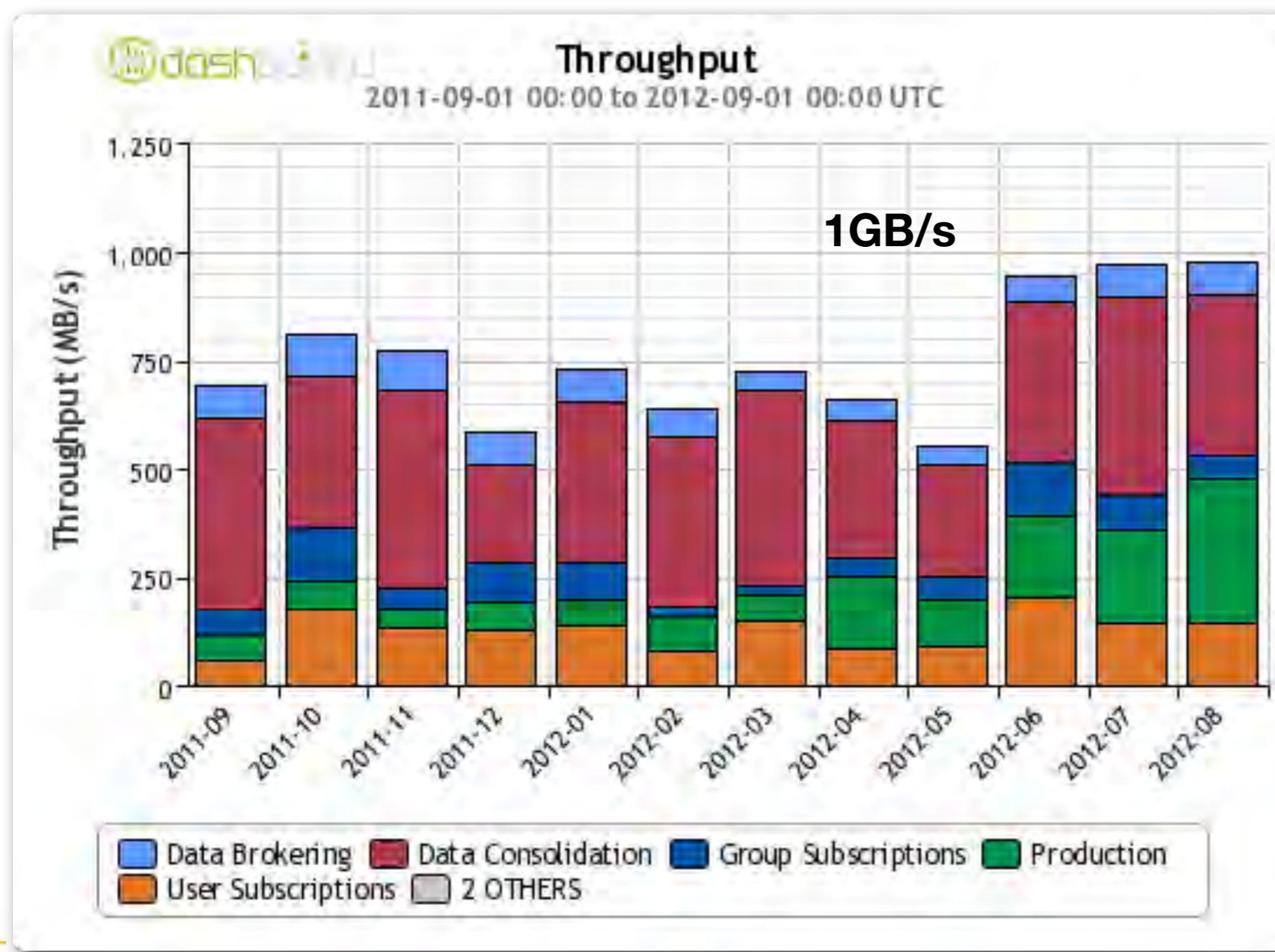


**78% to T1s**



**T1 → T1**

**23,966 TB**



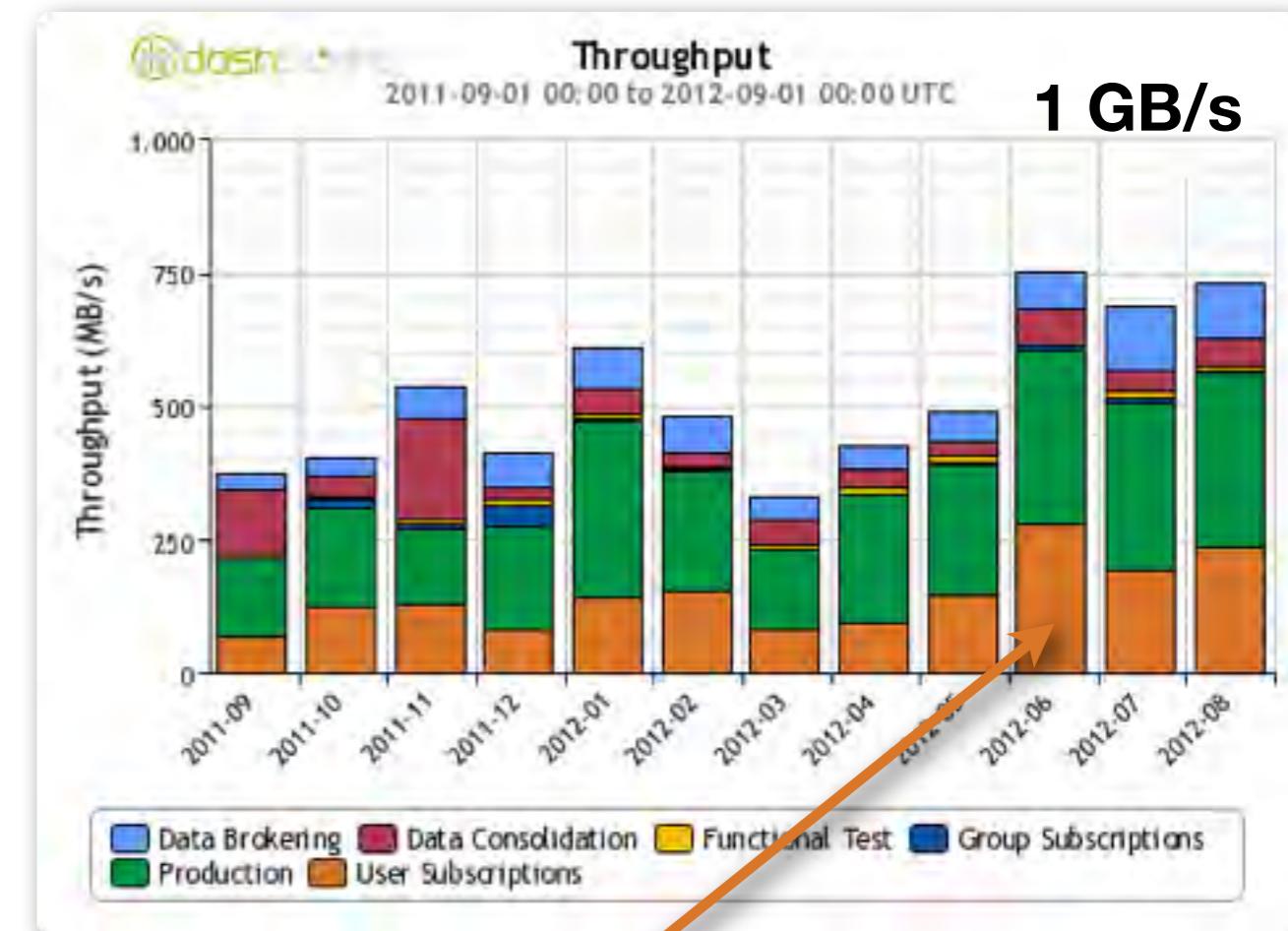
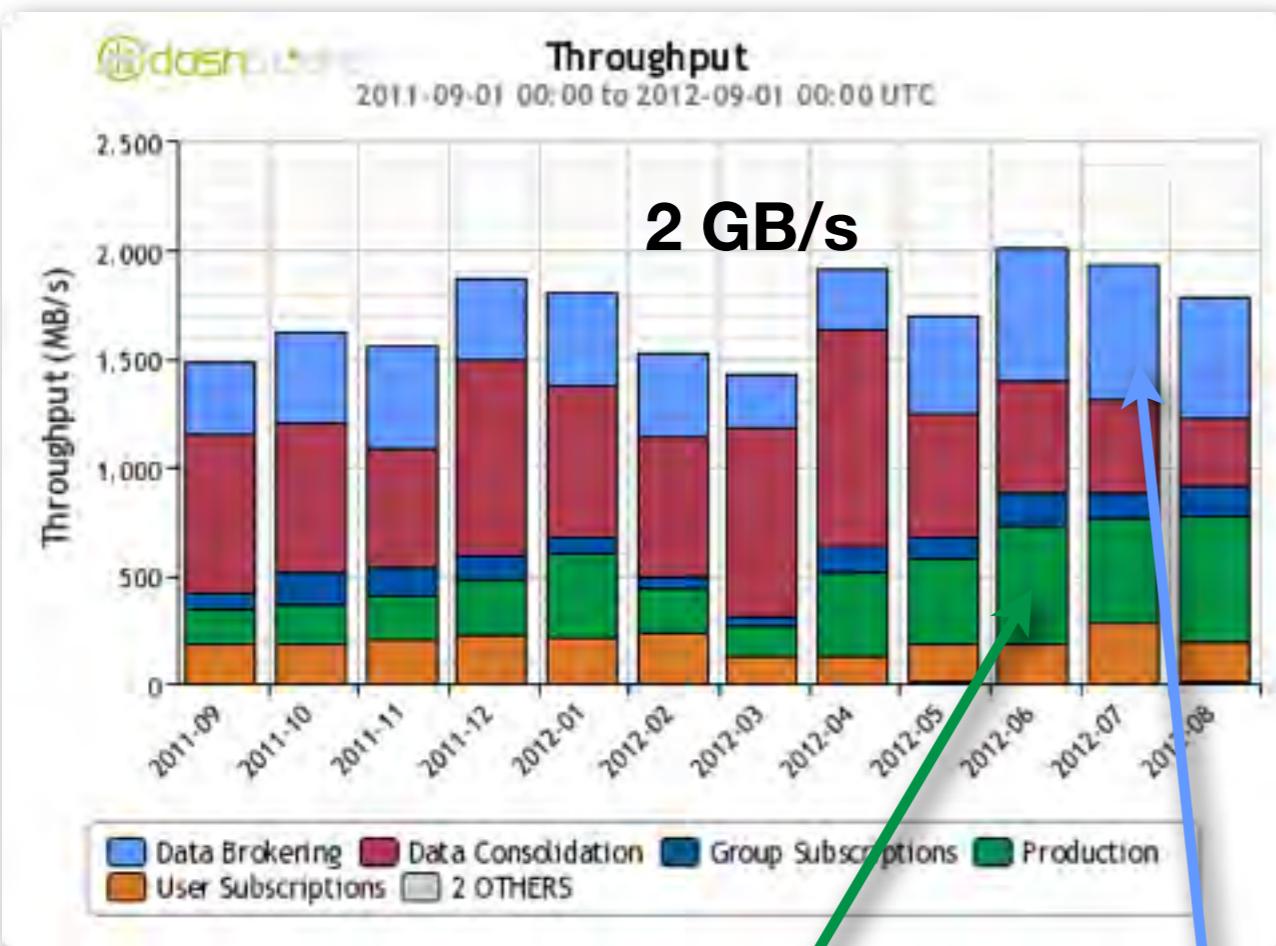
*Popularity based  
Pre-placement  
Group data  
Cross-cloud  
MC production  
User requests*

**T1 → T2**

**T2 → T1**

**54,491 TB**

**16,487 TB**



**Reconstruction in T2s**

**Dynamic data placement > pre-defined**

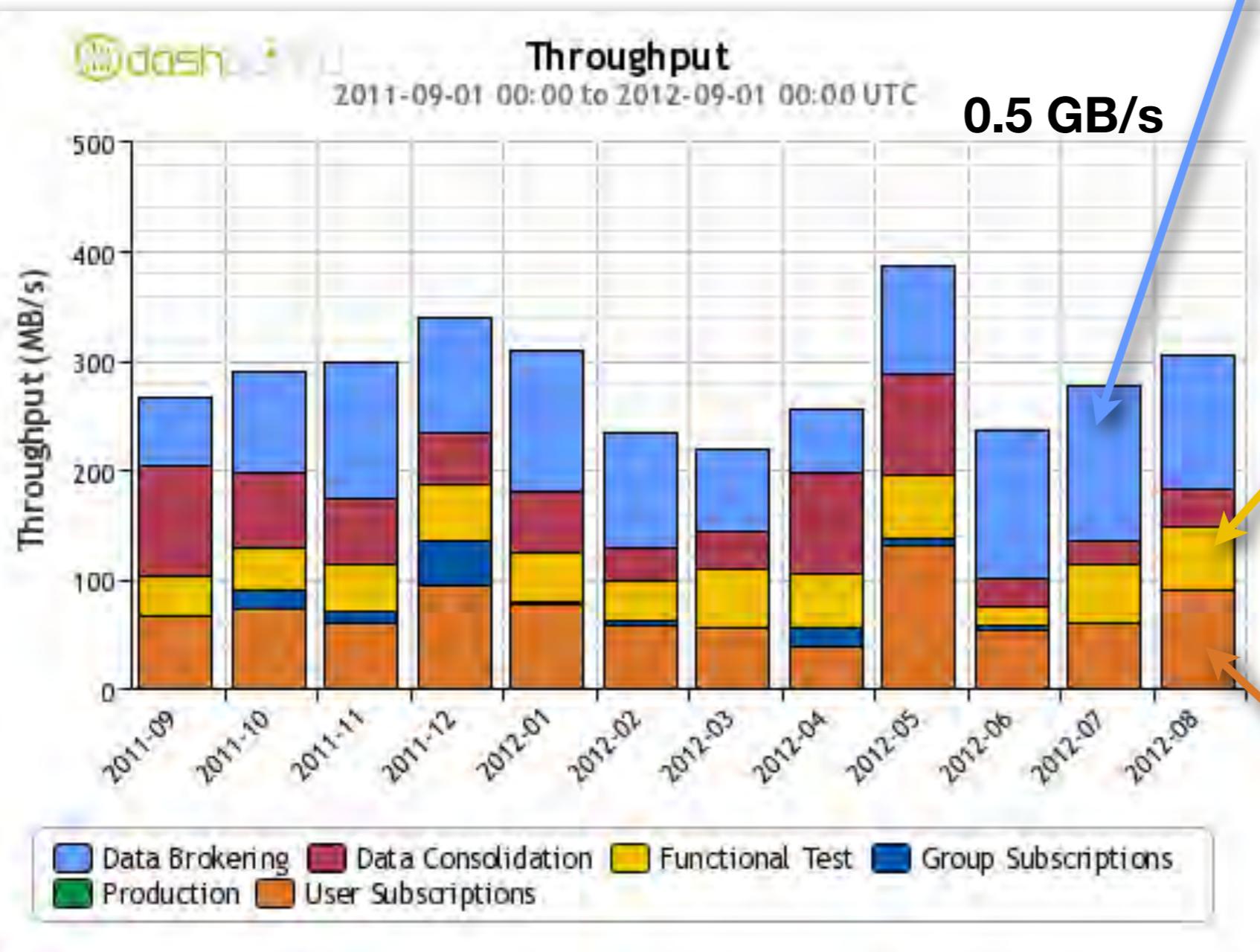
**User subscriptions!**

*Not in original computing model*

**T2→T2**

**Dynamic data placement > pre-defined**

**9,050 TB**



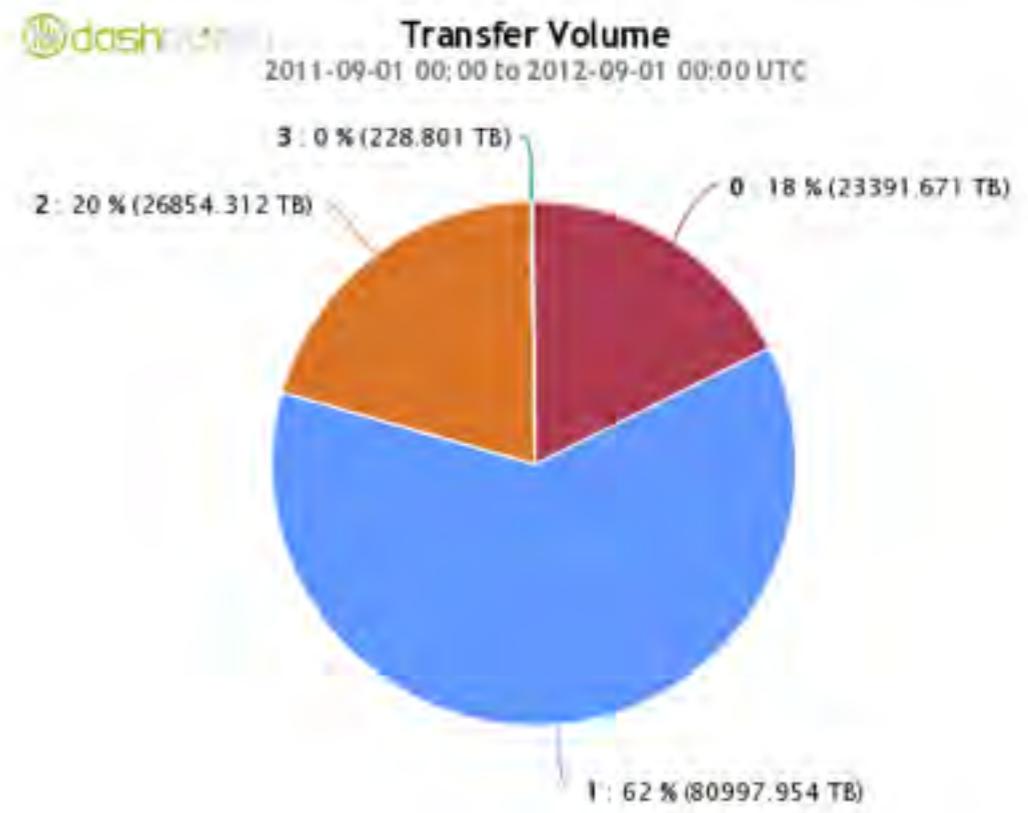
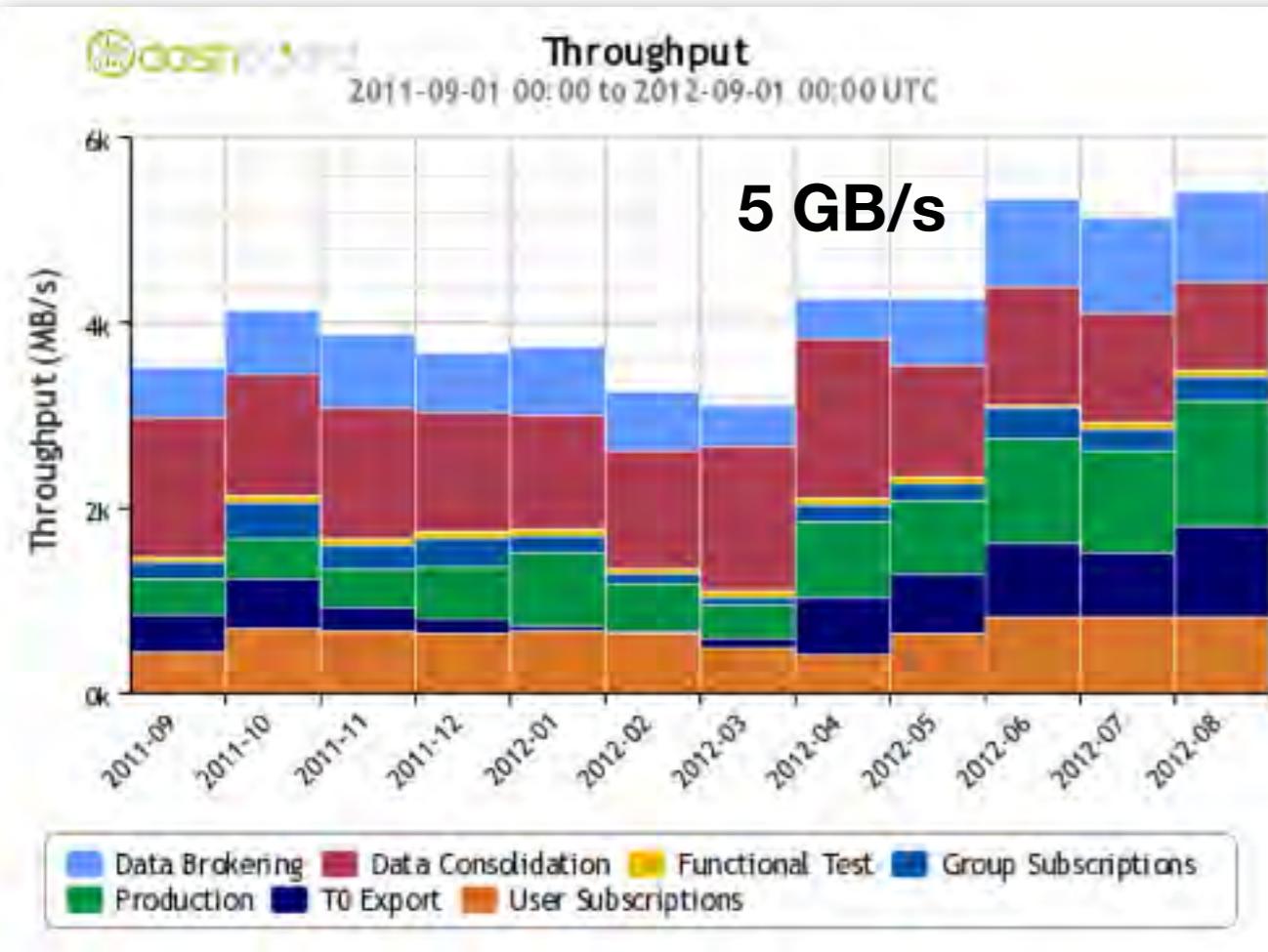
**Network mesh tests**

**User subscriptions  
Group data at some T2s  
+ Outputs of analysis**

Data volume: T1s 60% of sources

# ALL together

131,473 TB

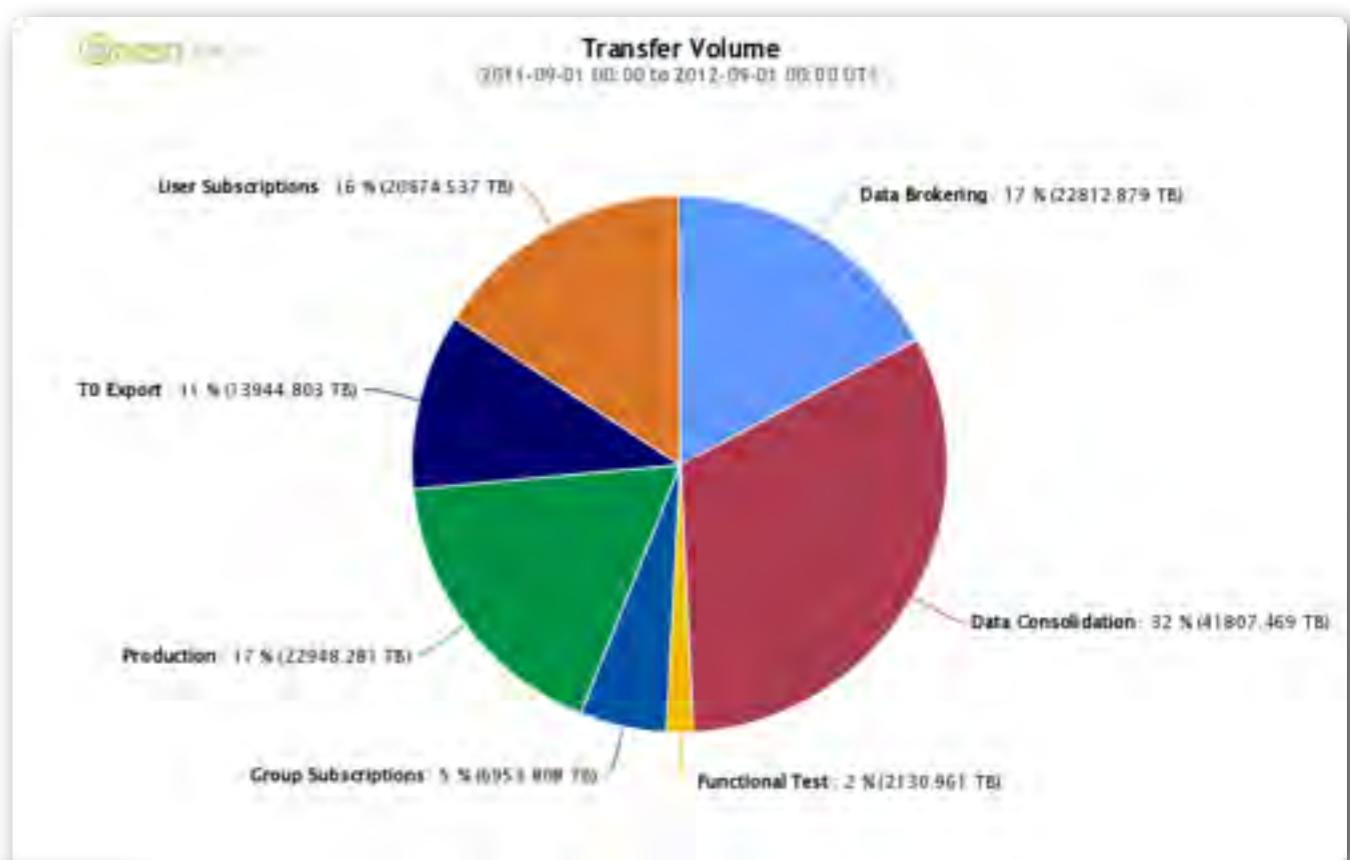


Activity: T2s 40% of sources

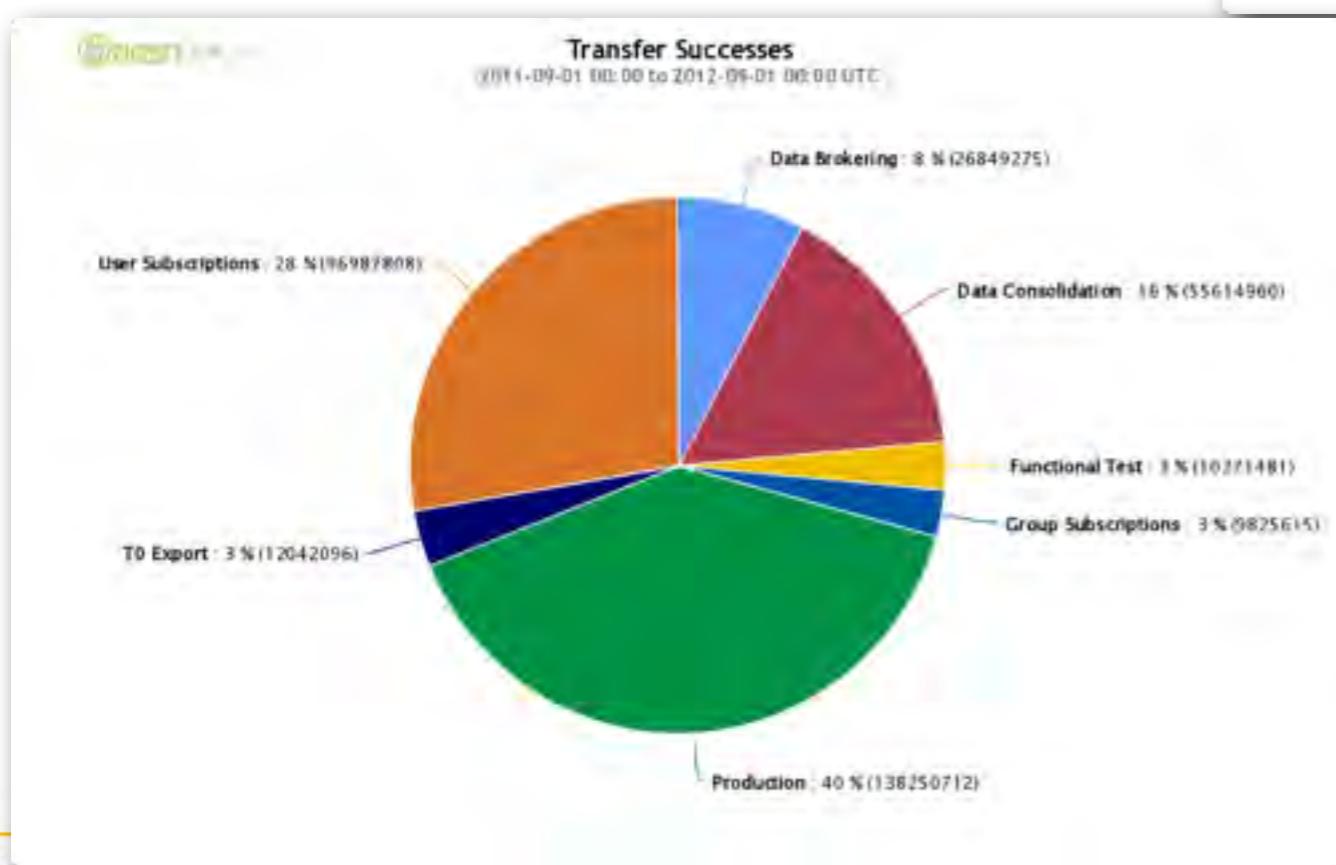


# ALL together

Data volume: pre-placement ~2 times  
dynamic placement room for improvements

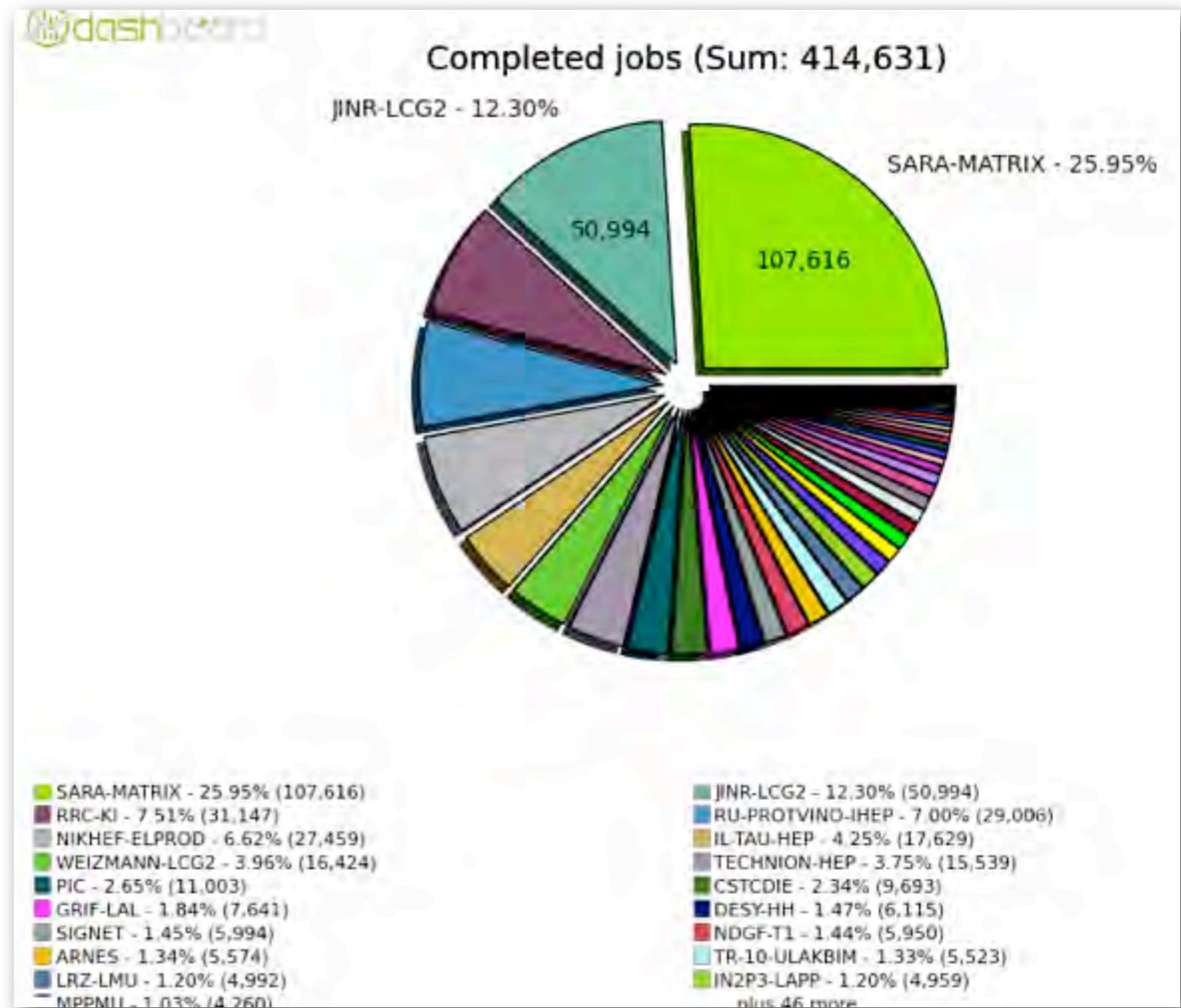


Activity: users ~30% of transfers



# Cross-cloud MC production

- The ‘easy’ part
- No need to be a T2D
- Only connection to remote T1 needed
- Example : NL cloud
  - 65 ! sites contributing



# The French cloud

---

- The most ‘exploded’ cloud of ATLAS
- 4 Romanians sites at the far end of GEANT
- 2 sites in far east Beijing & Tokyo connected to CCIN2P3 via different paths

T1 : Lyon

T2s : 14 sites

- Annecy
- Clermont
- Grenoble
- Grif (3 sites)
- Lyon
- Marseille
- Beijing
- Romania x4
- Tokyo

● ΤΟΚΙΟ

● ΒΟΜΒΑΣΙΣ X4

● ΡΕΙΓΚΙΝ

# The “French” cloud

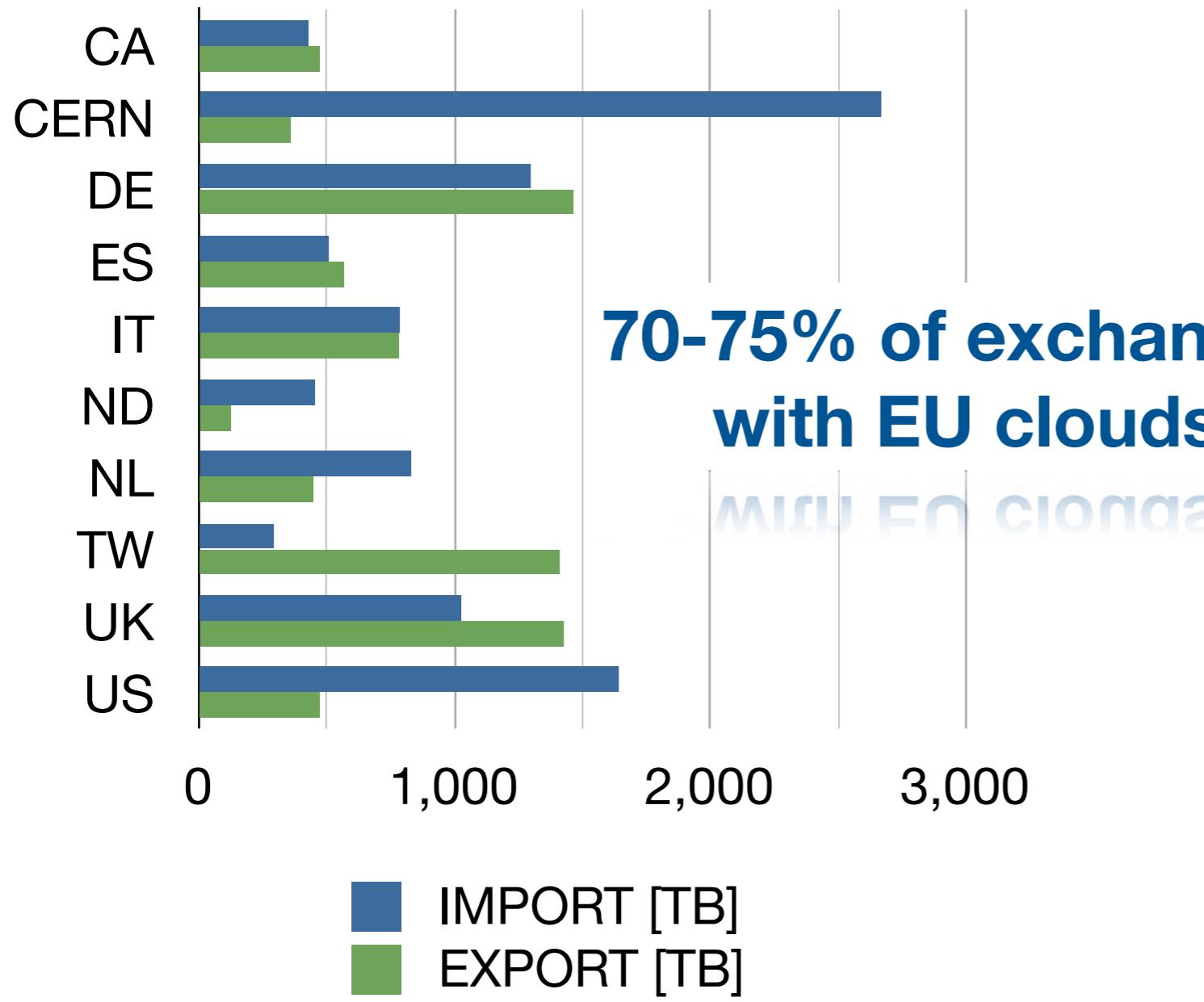


Data SIO, NOAA, U.S. Navy, NGA, GEBCO  
© 2012 Cnes/Spot Image  
Image © 2012 TerraMetrics

©2010 Google

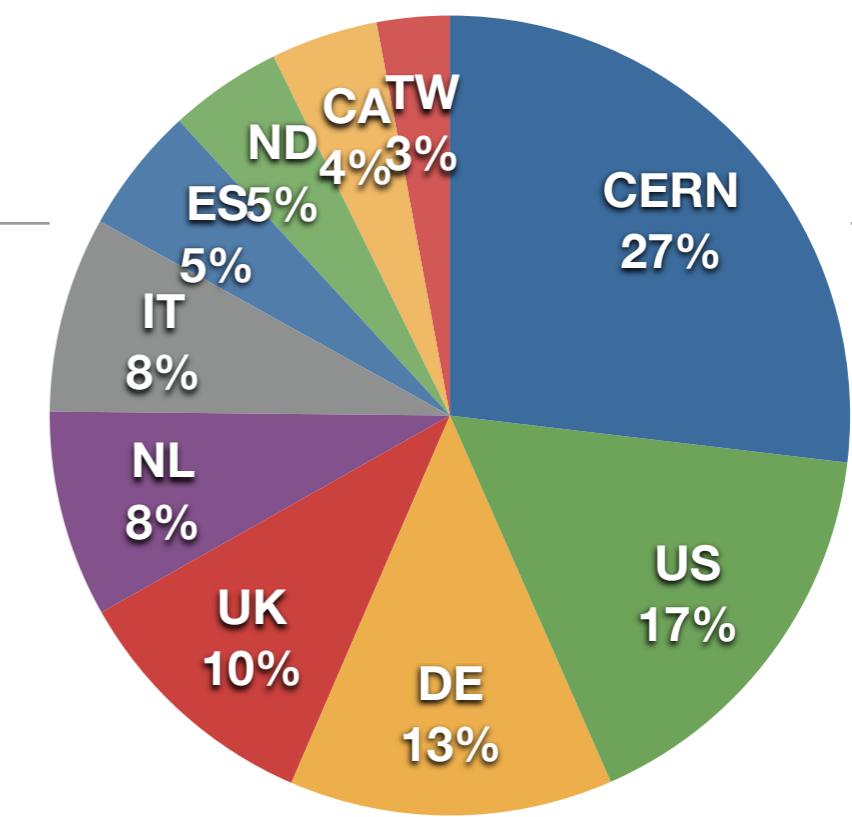
# Data exchanges for FR-cloud

## French cloud exchanges

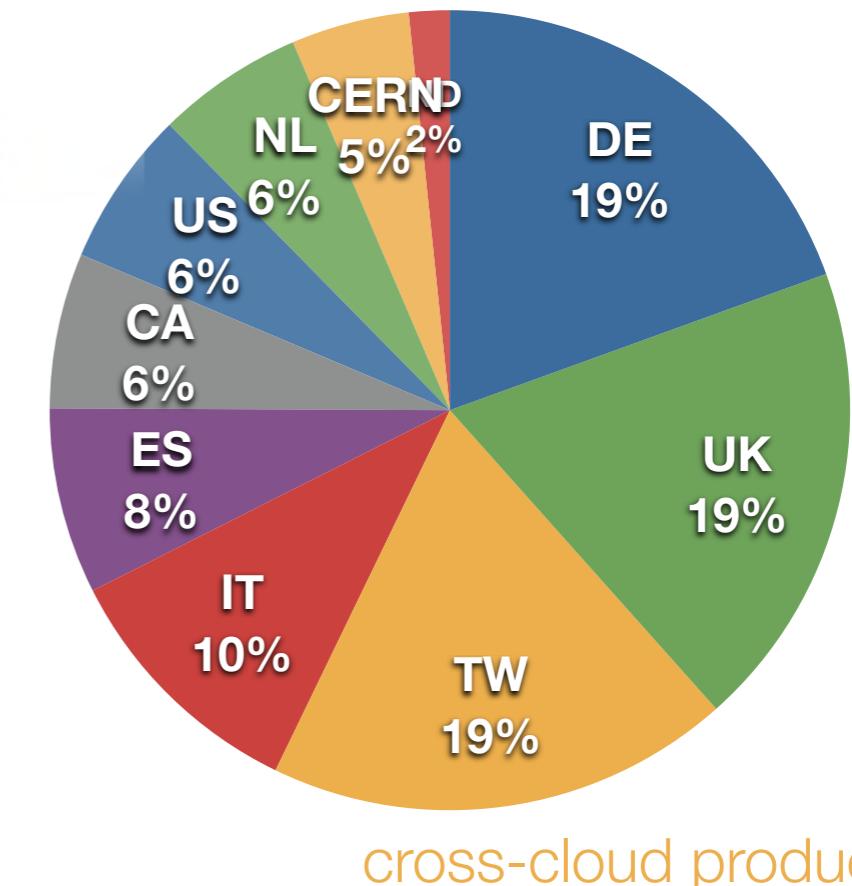


70-75% of exchanges  
with EU clouds

Imports : 9,9 PB



Exports : 7,5 PB



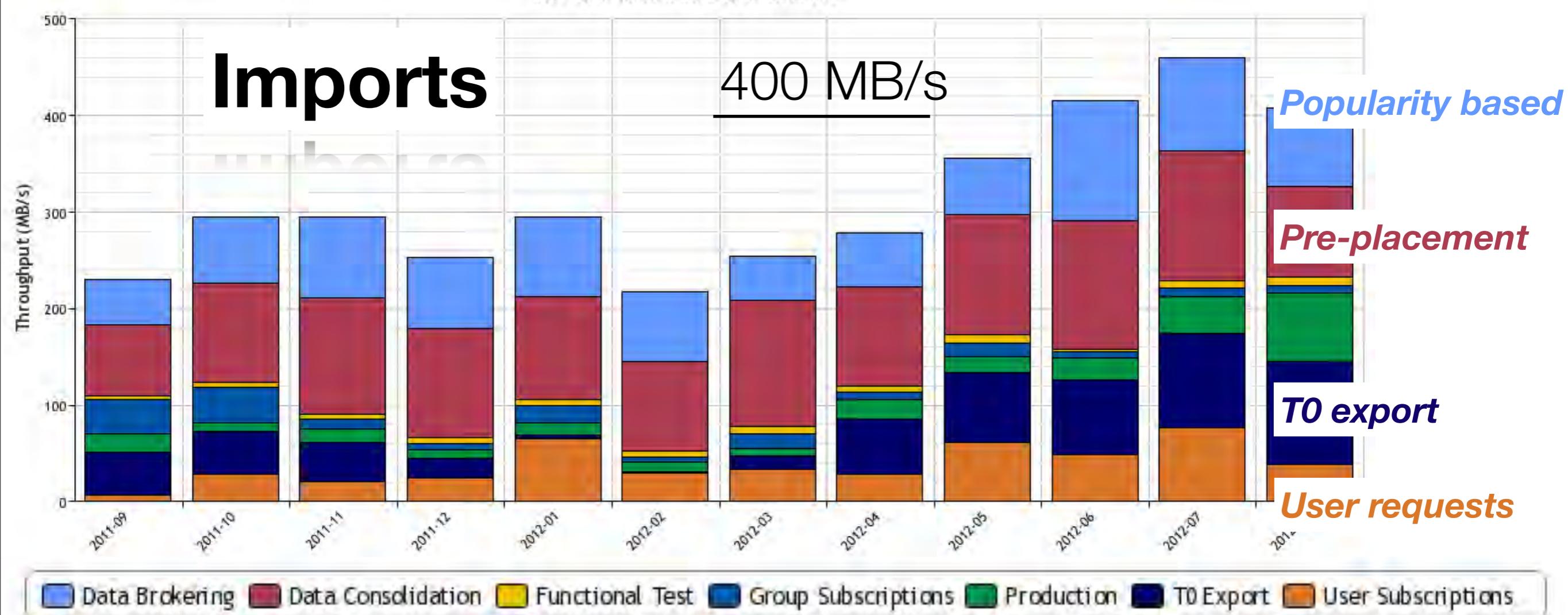
cross-cloud production

[Sep. 2011 - Sep. - 2012]

Rencontre LCG-France, SUBATECH Nantes, septembre 2012

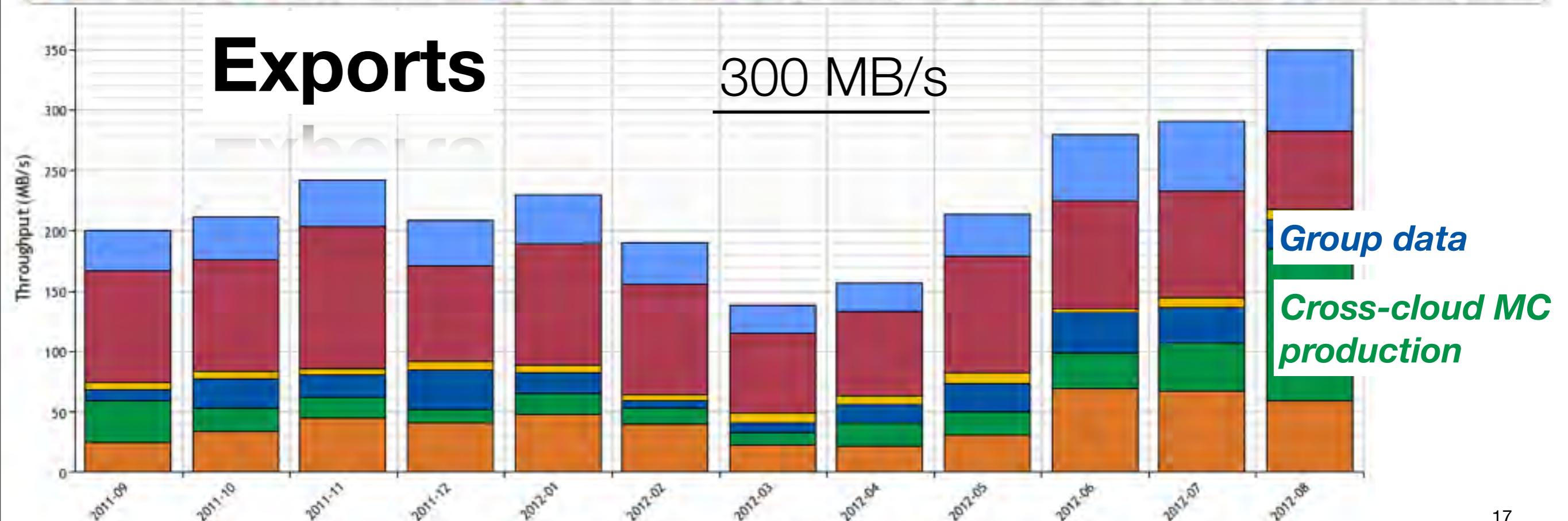
# Imports

400 MB/s

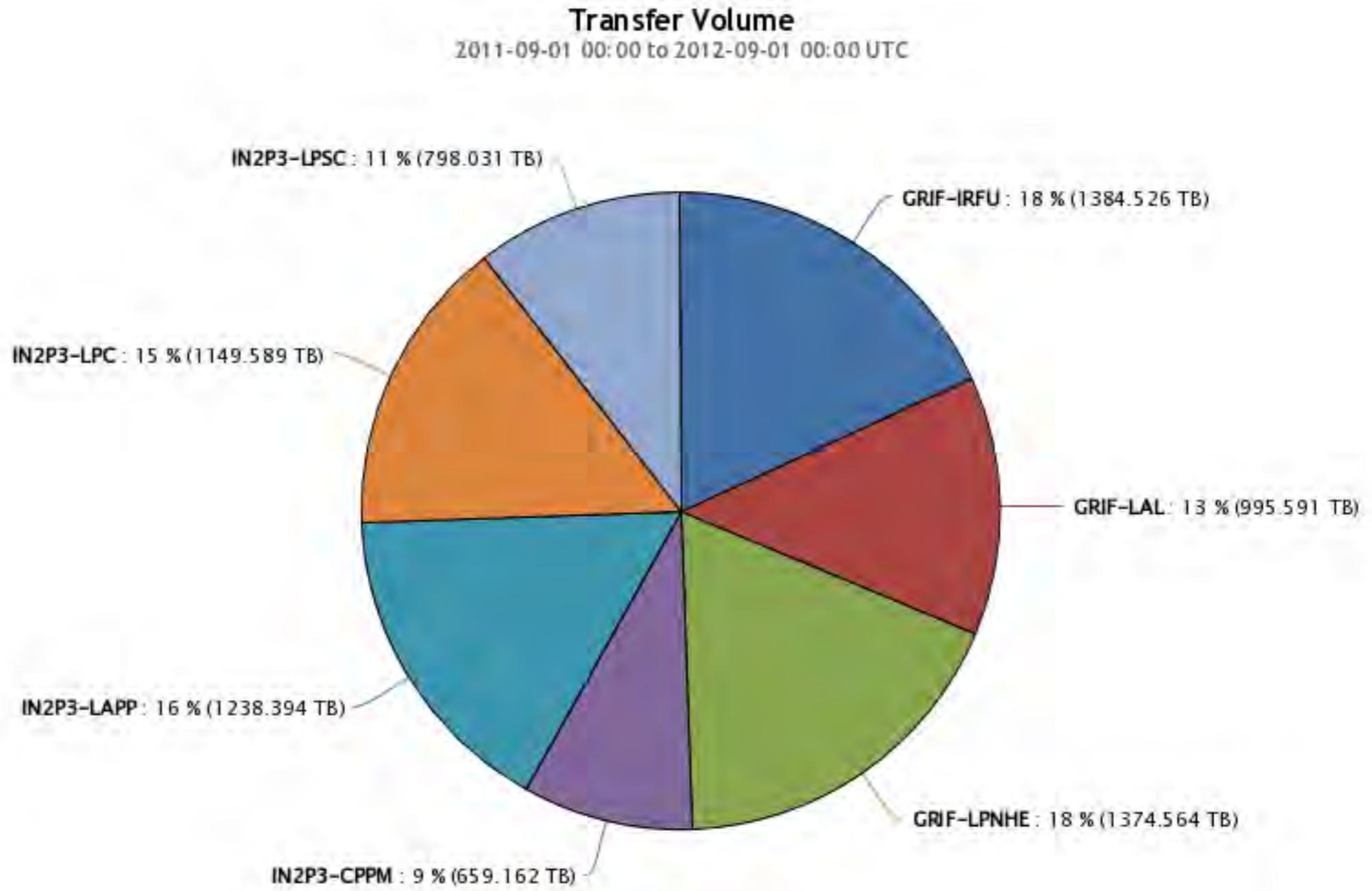


# Exports

300 MB/s



# Data volume transferred to French T2s

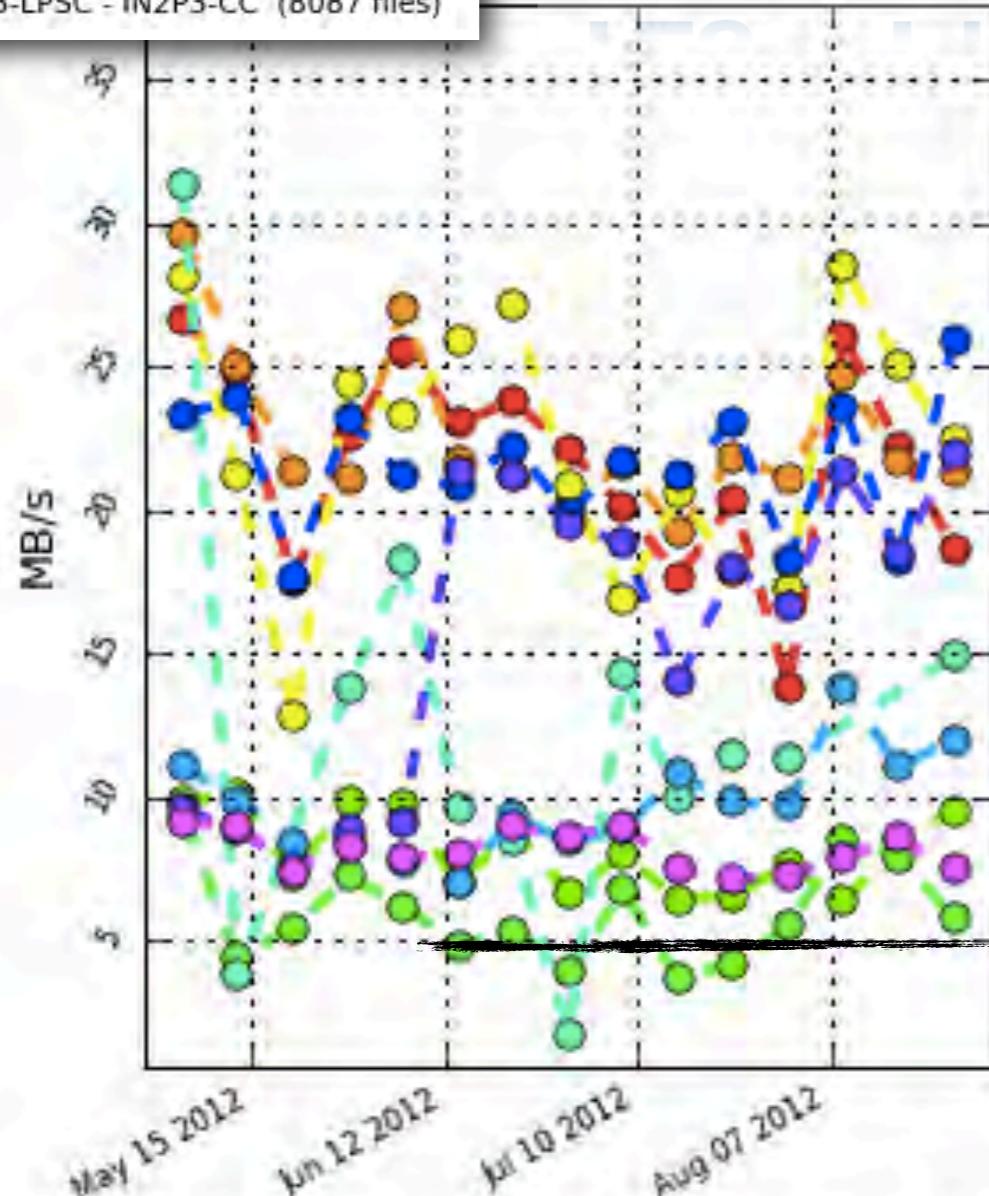


not proportional to number physicist nor CPUs

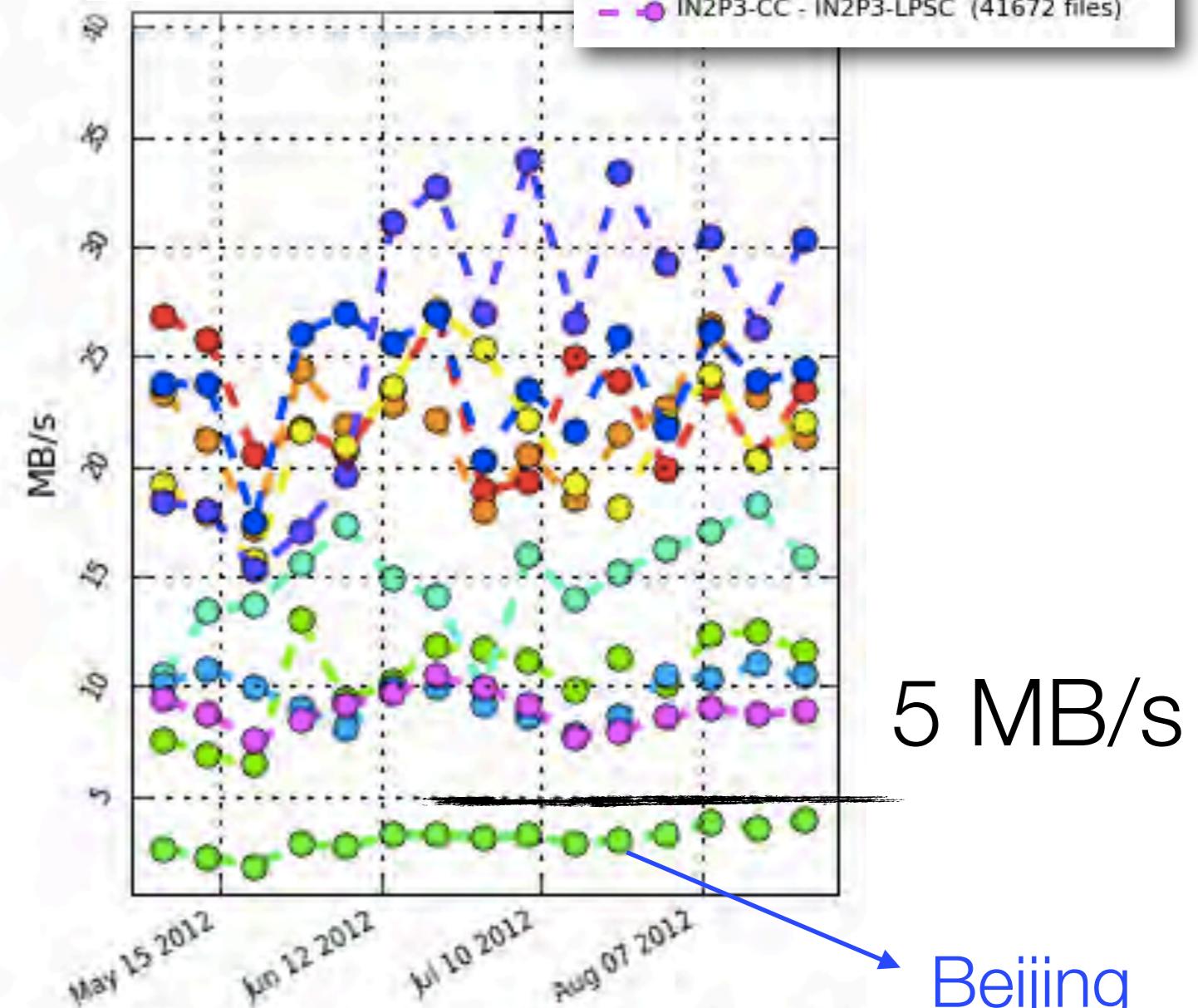
# Connectivity within French cloud (ATLAS sonar)

- GRIF-IRFU - IN2P3-CC (23264 files)
- GRIF-LAL - IN2P3-CC (15696 files)
- GRIF-LPNHE - IN2P3-CC (20109 files)
- TOKYO-LCG2 - IN2P3-CC (30237 files)
- BEIJING-LCG2 - IN2P3-CC (6715 files)
- RO-02-NIPNE - IN2P3-CC (1102 files)
- IN2P3-LAPP - IN2P3-CC (7179 files)
- IN2P3-LPC - IN2P3-CC (11170 files)
- IN2P3-CPPM - IN2P3-CC (12098 files)
- IN2P3-LPSC - IN2P3-CC (8087 files)

**T2s → T1**



**T1 → T2s**



5 MB/s

Beijing

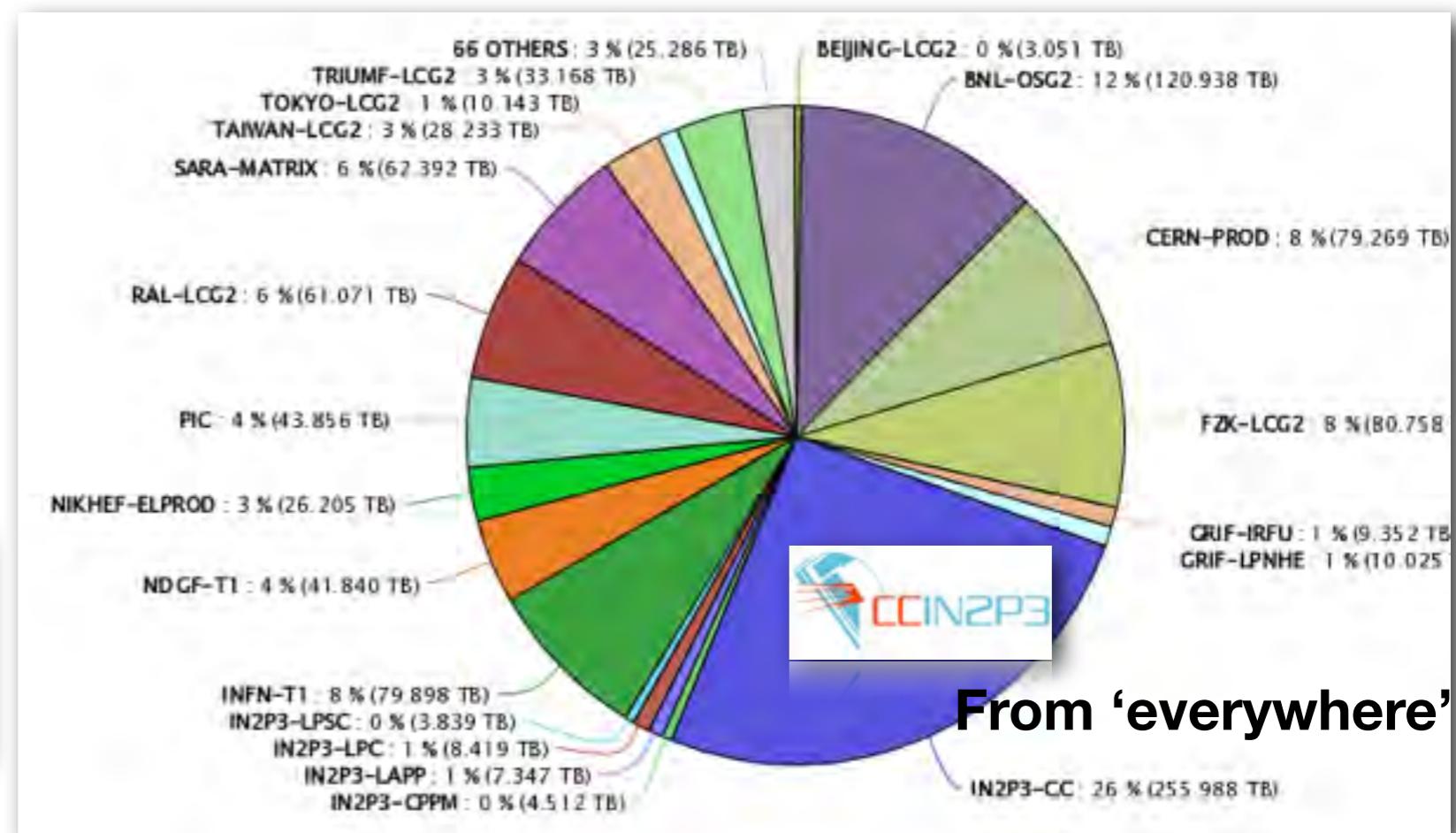
# Origin of data transferred to 2 GRIF sites

LAL : T2D  
connected to LHCONE

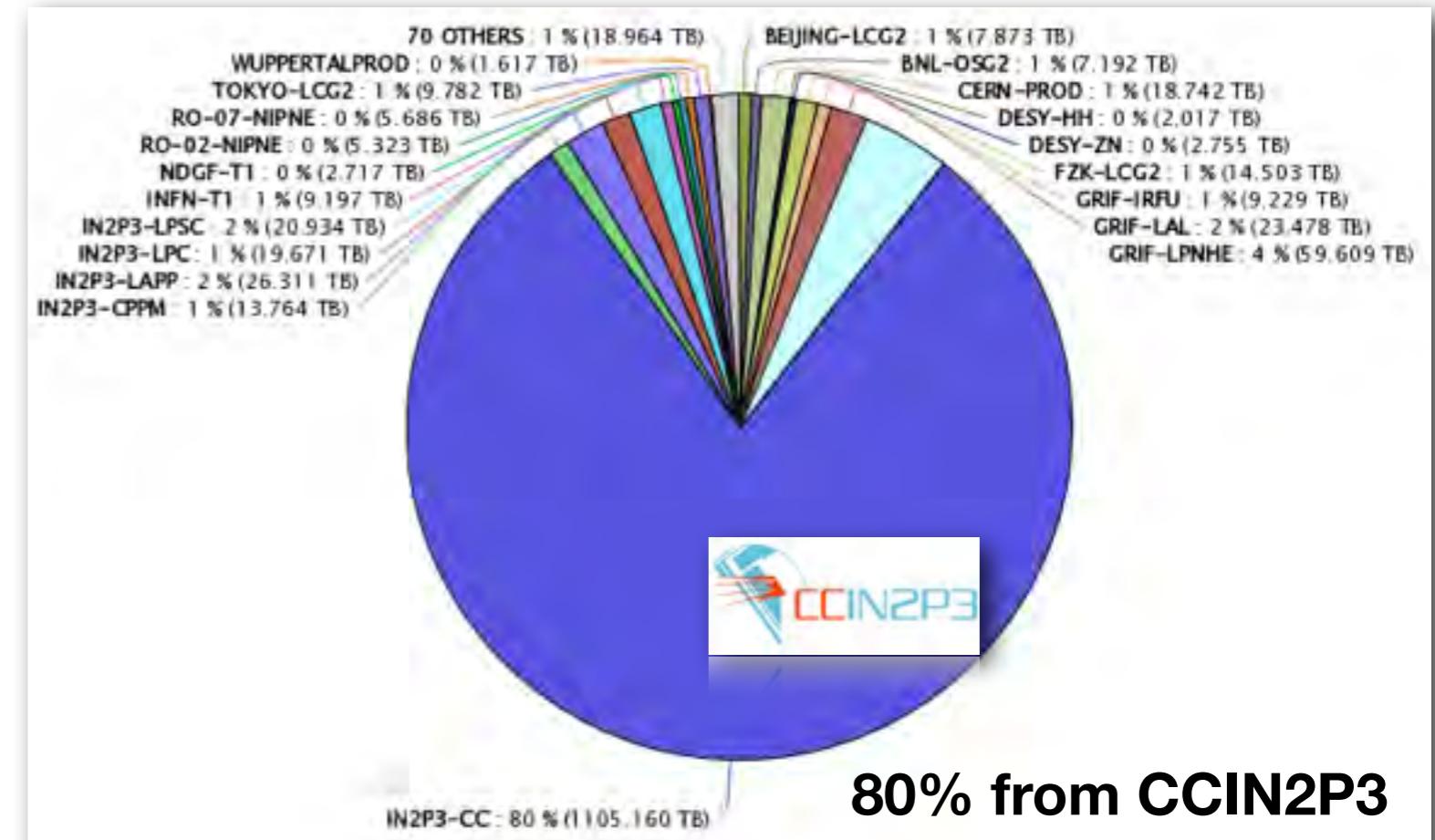
connected to LHCONE

IRFU : T2D  
connected to LHCONE

connected to LHCONE



From 'everywhere'



80% from CCIN2P3

# User & group analysis

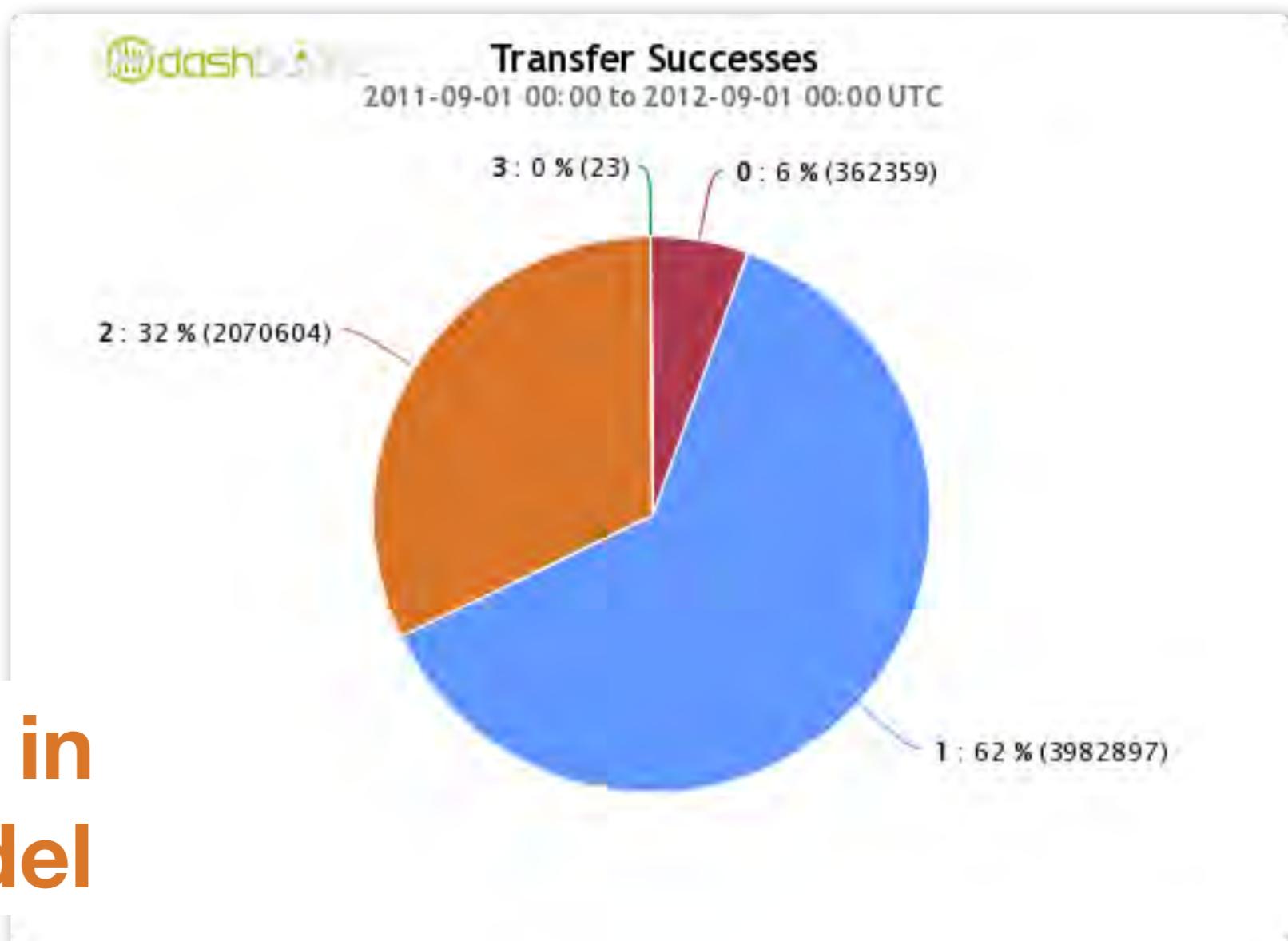
---

- Most of users run final analysis at their local site : delivery of data or analysis job outputs to users **very sensitive** (the last transferred file determines the efficiency)
- 2 ways to get (reduced format) data
  - The majority : Let PanDA decide where to run the jobs ; where data are (in most of the cases). Outputs stored on SCRATCHDISK (buffer area) at remote sites have to be shipped back to user local site
  - Get limited volume of data at local site to run locally analysis
- Both imply data transfers from remote site to local site
  - Direct transfers for T2Ds
  - Through T1 for other T2s

# User + Group transfers to FR-cloud sites

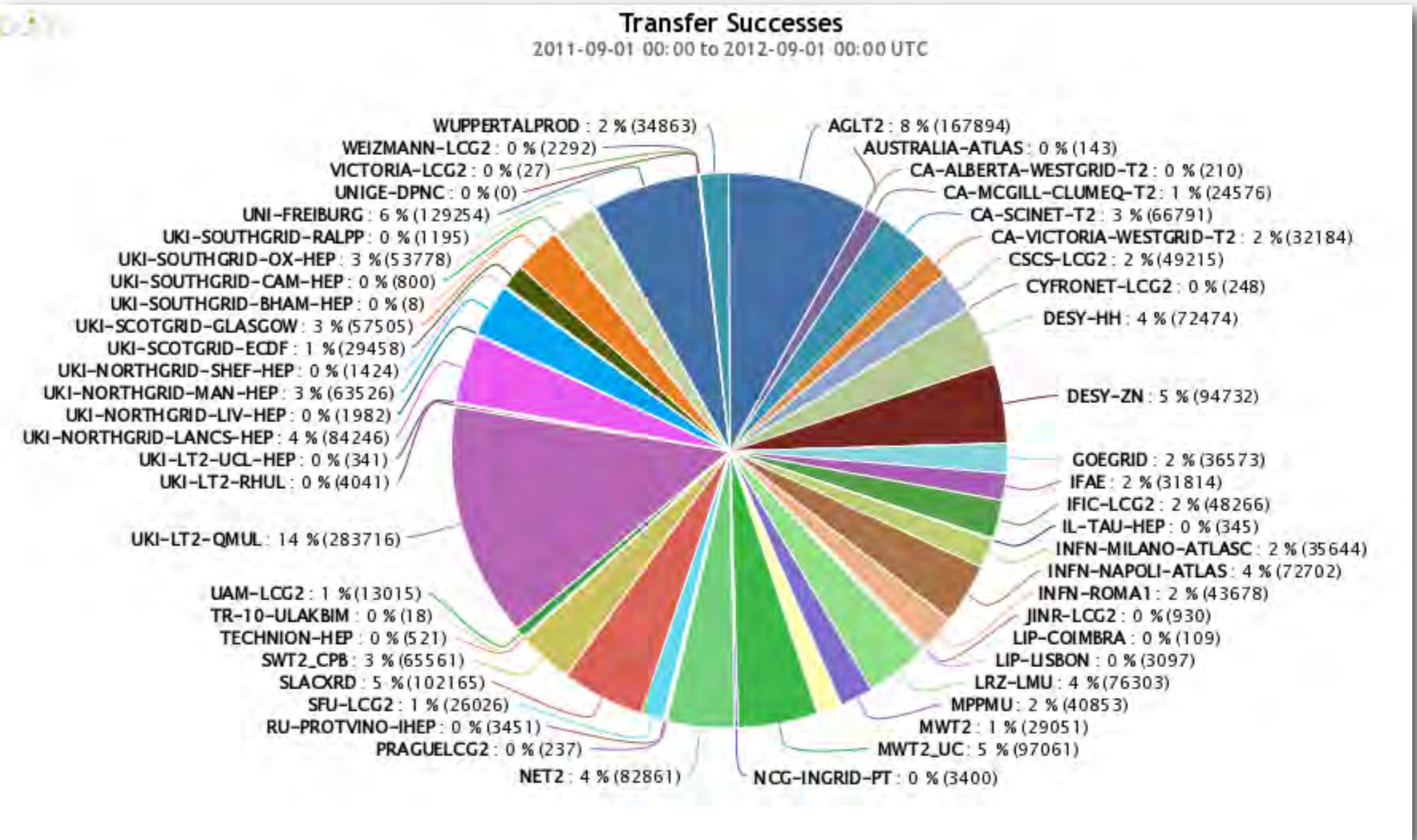
$\frac{1}{3}$  of data transfers  
(not data volume)  
used for final  
analysis from **T2**  
**sites outside FR-**  
cloud

1,828 TB

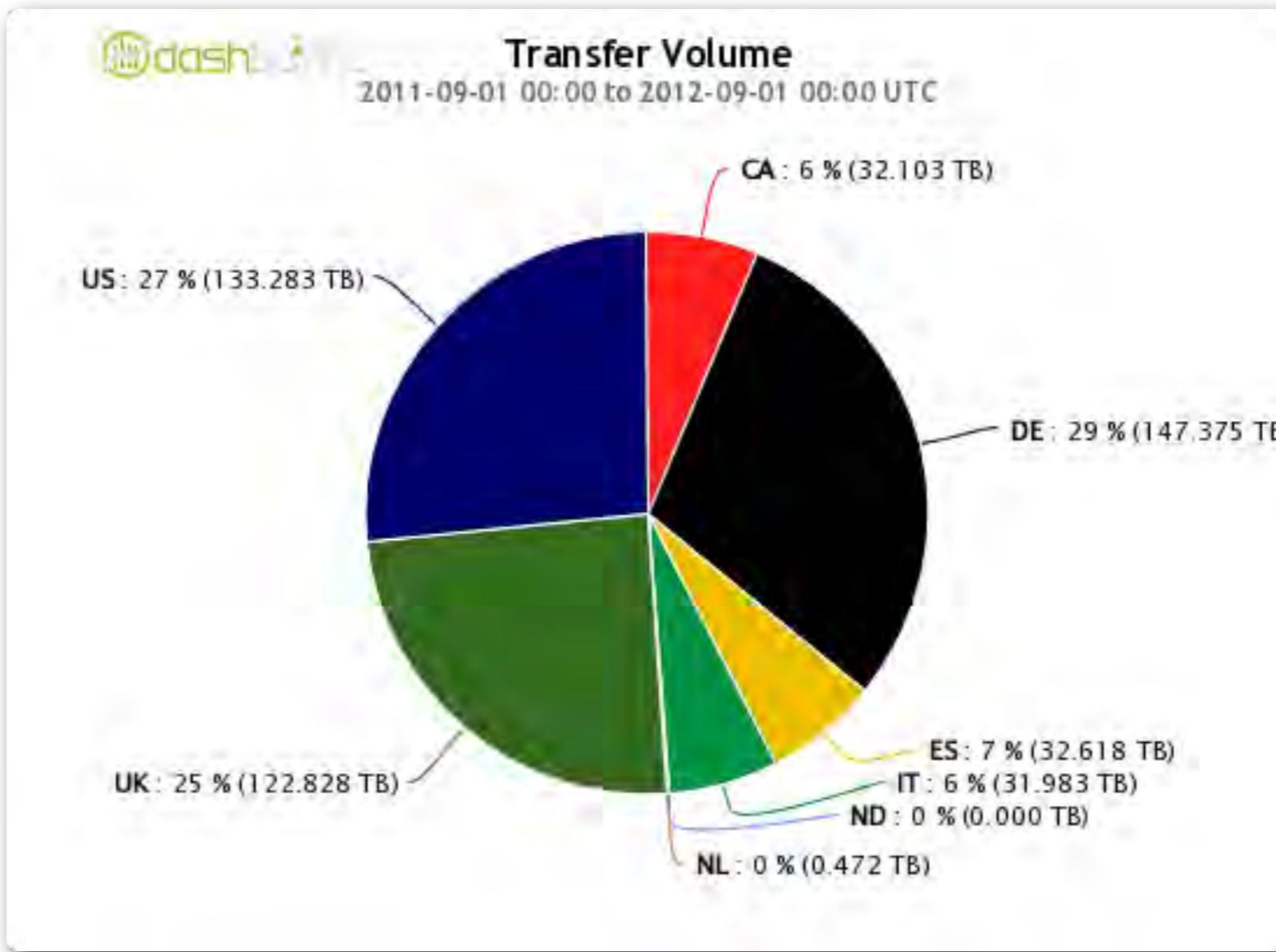


**Not allowed in  
original model**

# Data come from 52 T2 sites



$\frac{1}{3}$  from non EU T2s



$\frac{1}{4}$  from UK (not on LHCONE)

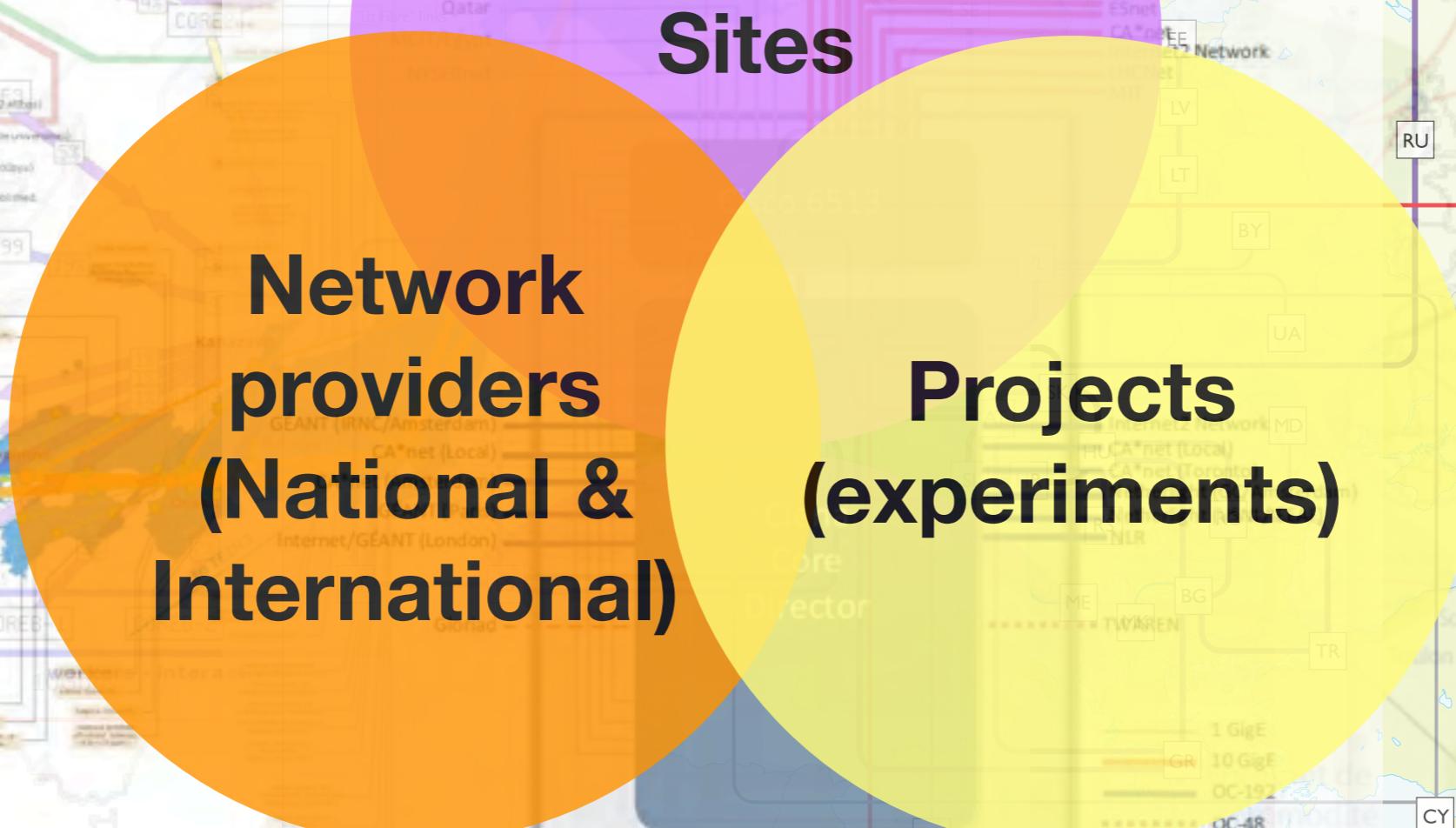
# Issues with distant T2s

# Beijing

---

- **Beijing :**
  - Connected to Europe via GEANT/TEIN3 (except CERN : GLORIAD/KREONET)
  - RTT ~190 ms
- **Tokyo :**
  - Connected to Europe via GEANT/MANLAN/SINET4
  - RTT ~ 300 ms
- Several network operators on the path (Nationals, GEANT, ...)

# The difficulty to solve network issues



**Various approaches and complementary tools needed**

# perfSonar dashboard of FR-cloud

Being expanded as sites install perfSonar

**RACF**  
Grid Group

**The Production Instance of perfSONAR Dashboard**

Status as of: Thu Sep 13 13:53:10 EDT 2012

**Cloud LHC-FR**

Sites of LHC-FR cloud

CC-IN2P3	Tokyo	GRIF-LAL	Beijing	GRIF/LPNHE	RO-07
RO-02					

**LHC-FR Throughput**

	---	0	1	2	3	4	5	6
0:Beijing (perfsonar.ihep.ac.cn)	---	0.15	0.00	0.00	0.00	0.70		
1:CC-IN2P3 (ccperfsonar-lhcopn.in2p3.fr)	0.27	---	0.00	0.00	0.00	0.43		
2:GRIF-LAL (psonar2.lal.in2p3.fr)	0.00	0.80	---	0.00	0.00	0.23		
3:GRIF/LPNHE (lpnhe-psb.in2p3.fr)	0.00	0.00	0.91	---	0.18	0.09	0.25	
4:RO-02 (atrogr009.nipne.ro)	0.09	0.00	0.20	0.00	---	0.94	0.23	
5:RO-07 (perfsonar1.nipne.ro)	0.15	0.00	0.67	0.00	0.94	---	0.00	
6:Tokyo (perfsonar2.icepp.jp)	0.66	0.00	0.23	0.24	0.18	0.00	---	

**LHC-FR Packet Loss**

	---	0	1	2	3	4	5	6
0:Beijing (perfsonar2.ihep.ac.cn)	---	0.0	0.0	0.0	0.0	0.0	0.0	
1:CC-IN2P3 (ccperfsonar2-lhcopn.in2p3.fr)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2:GRIF-LAL (psonar1.lal.in2p3.fr)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3:GRIF/LPNHE (lpnhe-psl.in2p3.fr)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4:RO-02 (atrogr007.nipne.ro)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
5:RO-07 (perfsonar2.nipne.ro)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
6:Tokyo (perfsonar1.icepp.jp)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

**performance ps toolkit**

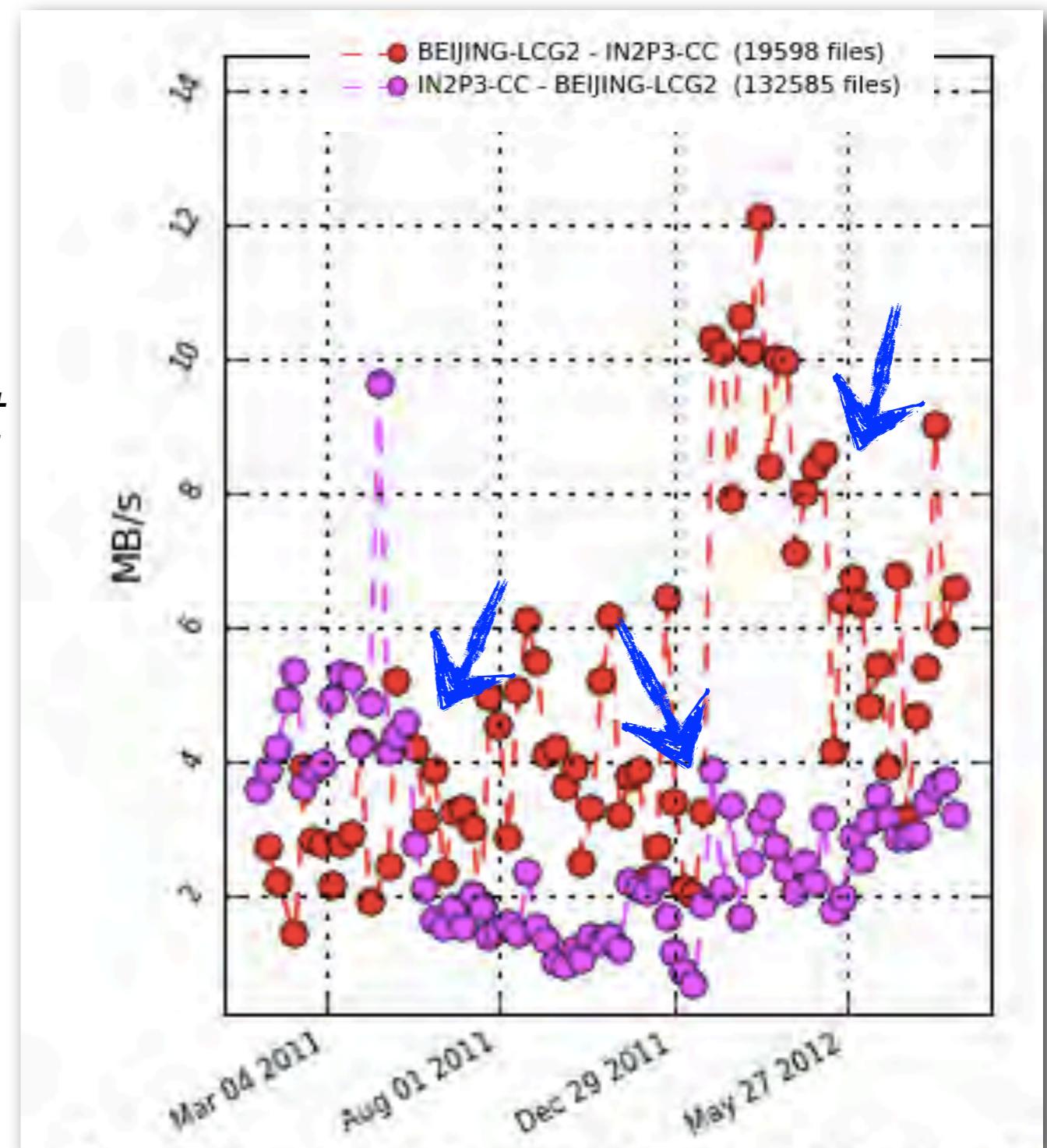
# ATLAS transfers to Beijing since beg. 2011

**Beijing → CCIN2P3  
CCIN2P3 → Beijing**

*Performances changed over last year*

- Asymmetry in transfer rate : why?
- Asymmetry reversed

**Each ‘event’ explained  
sometime after some delay...**



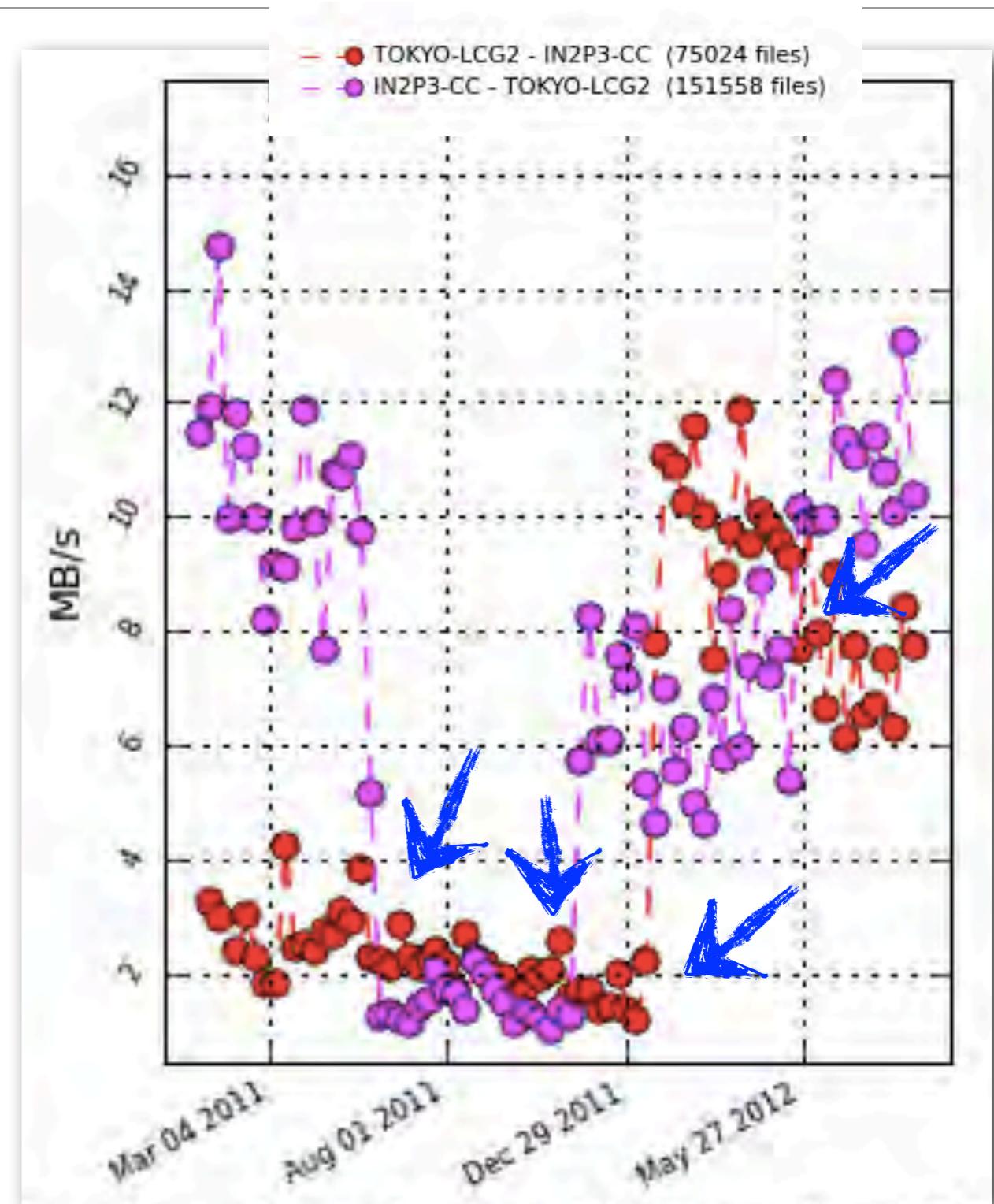
# ATLAS transfers to Tokyo since beg. 2011

**Tokyo → CCIN2P3  
CCIN2P3 → Tokyo**

*Performances changed over last year*

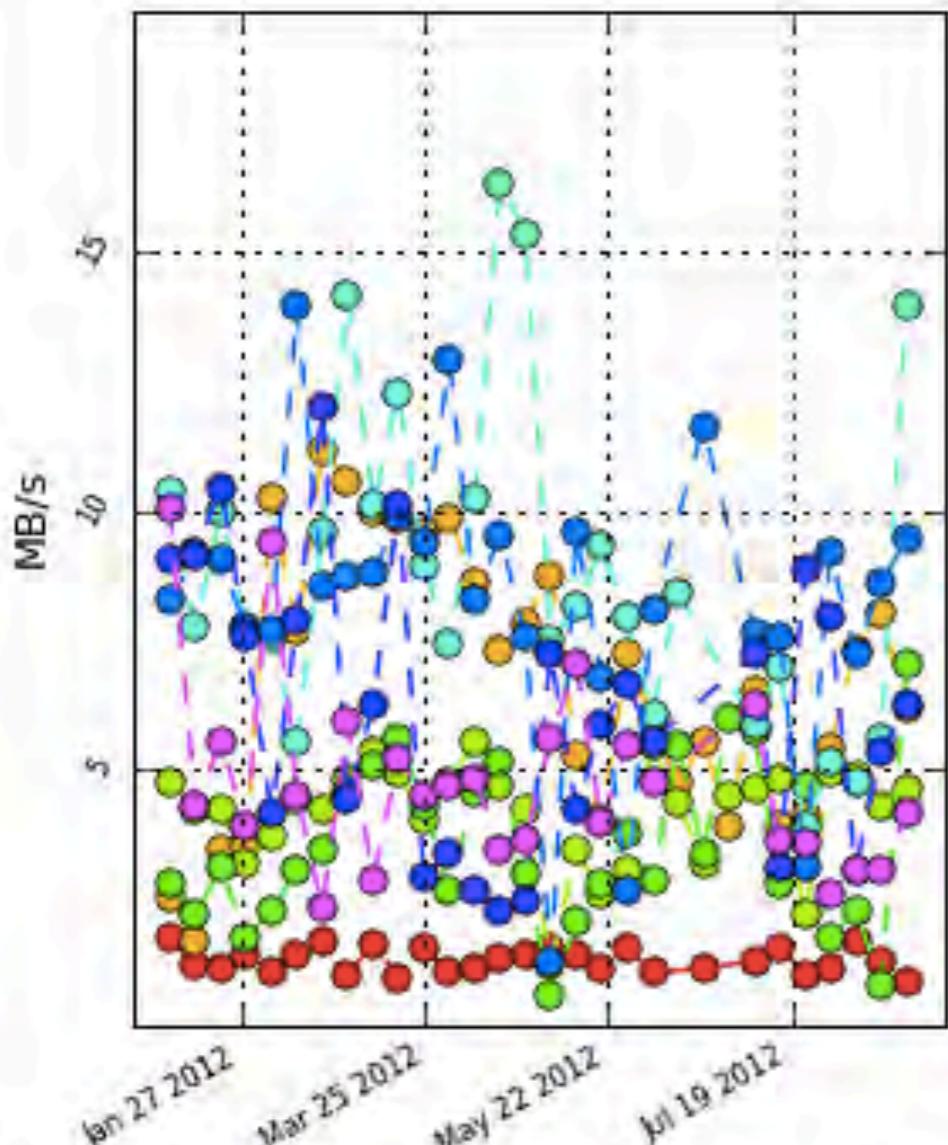
- Asymmetry in transfer rate : why?
- Asymmetry reversed

**Each ‘event’ explained  
sometime after some delay...**



# Beijing from/to EU T1s

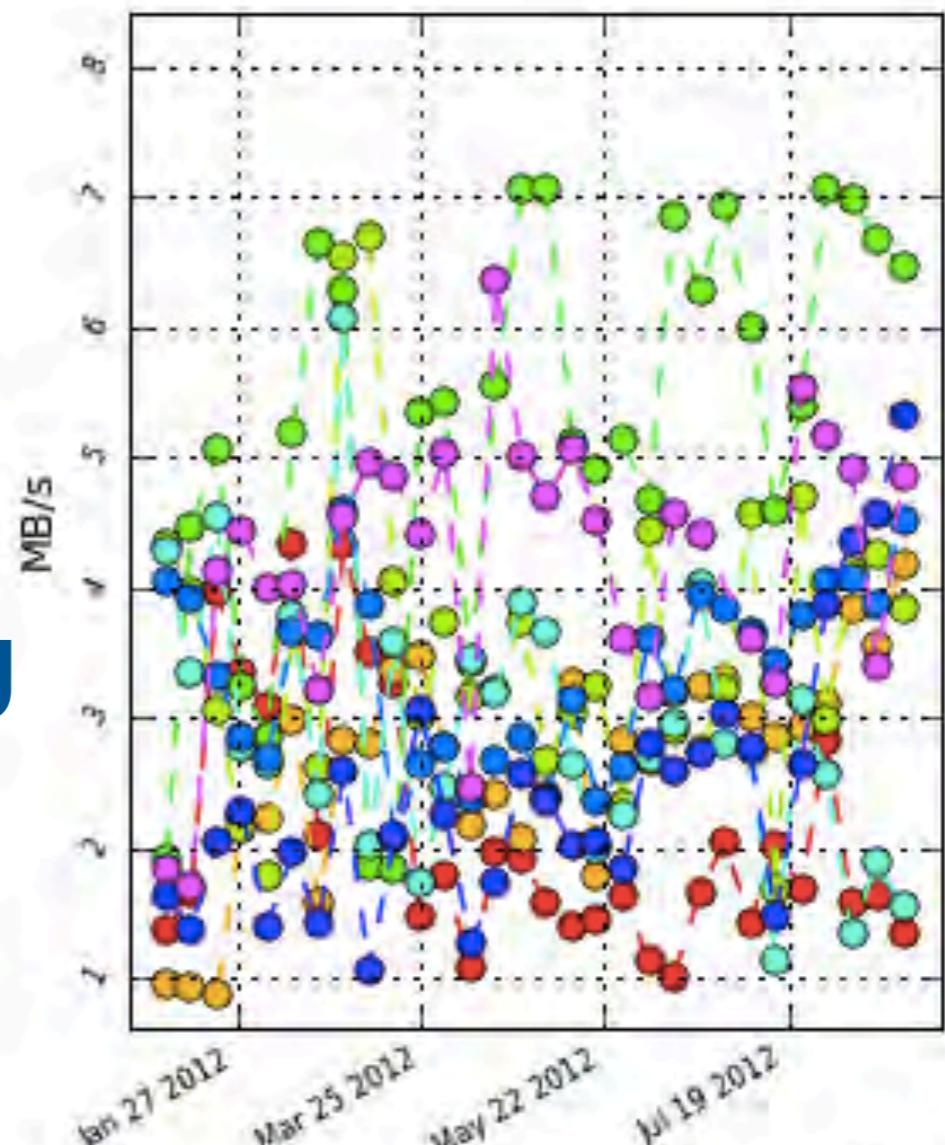
- BEIJING-LCG2 - FZK-LCG2 (1405 files)
- BEIJING-LCG2 - IN2P3-CC (9364 files)
- BEIJING-LCG2 - RAL-LCG2 (600 files)
- BEIJING-LCG2 - CERN-PROD (1128 files)
- BEIJING-LCG2 - SARA-MATRIX (1224 files)
- BEIJING-LCG2 - INFN-T1 (704 files)
- BEIJING-LCG2 - PIC (978 files)
- BEIJING-LCG2 - NDGF-T1 (530 files)



Each T1 is  
different

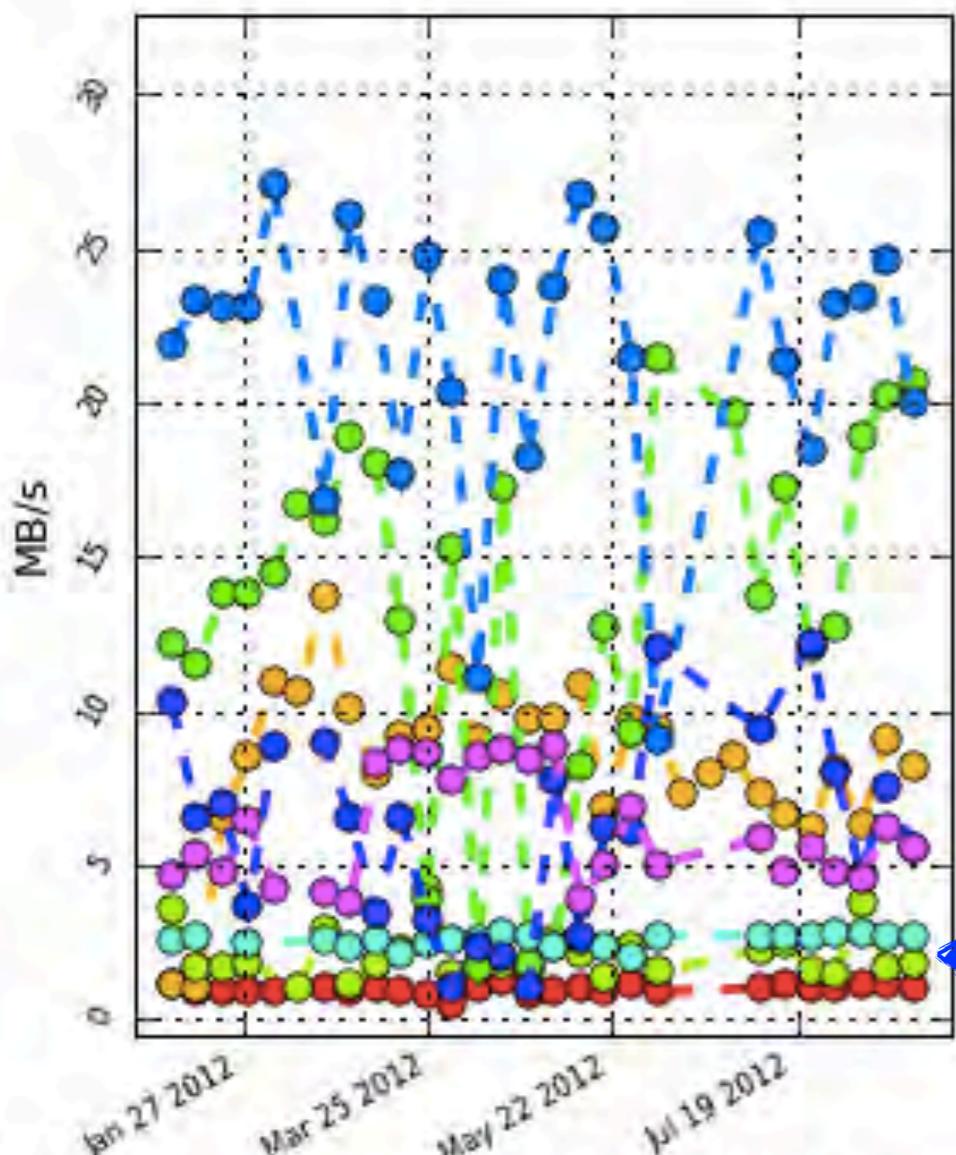
Beijing → EU  
better for  
most T1s

- FZK-LCG2 - BEIJING-LCG2 (6631 files)
- IN2P3-CC - BEIJING-LCG2 (73013 files)
- RAL-LCG2 - BEIJING-LCG2 (11370 files)
- CERN-PROD - BEIJING-LCG2 (13098 files)
- SARA-MATRIX - BEIJING-LCG2 (6296 files)
- INFN-T1 - BEIJING-LCG2 (12739 files)
- PIC - BEIJING-LCG2 (6962 files)
- NDGF-T1 - BEIJING-LCG2 (7877 files)



# Tokyo from/to EU T1s

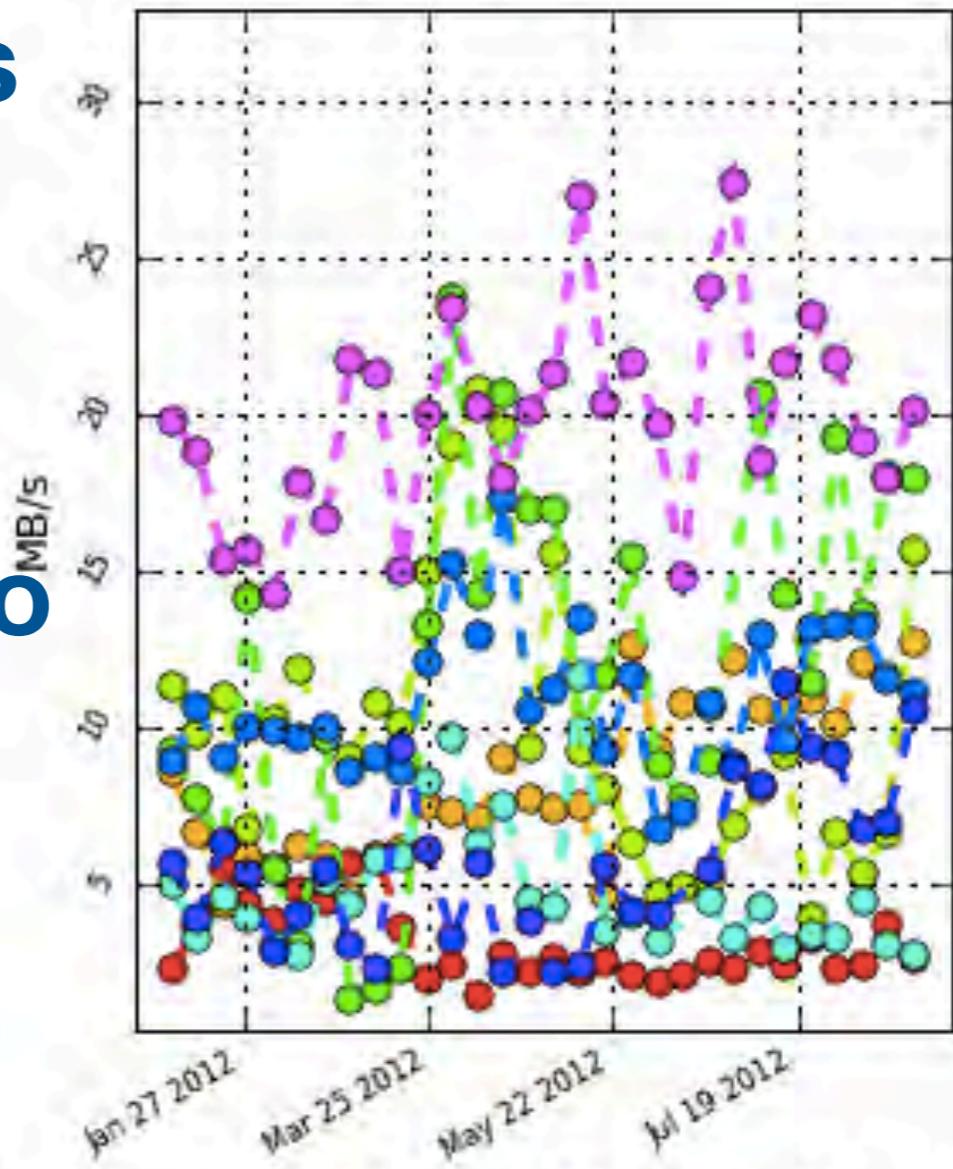
- TOKYO-LCG2 - FZK-LCG2 (150 files)
- TOKYO-LCG2 - IN2P3-CC (46972 files)
- TOKYO-LCG2 - RAL-LCG2 (147 files)
- TOKYO-LCG2 - CERN-PROD (187 files)
- TOKYO-LCG2 - SARA-MATRIX (140 files)
- TOKYO-LCG2 - INFN-T1 (1065 files)
- TOKYO-LCG2 - PIC (694 files)
- TOKYO-LCG2 - NDGF-T1 (152 files)



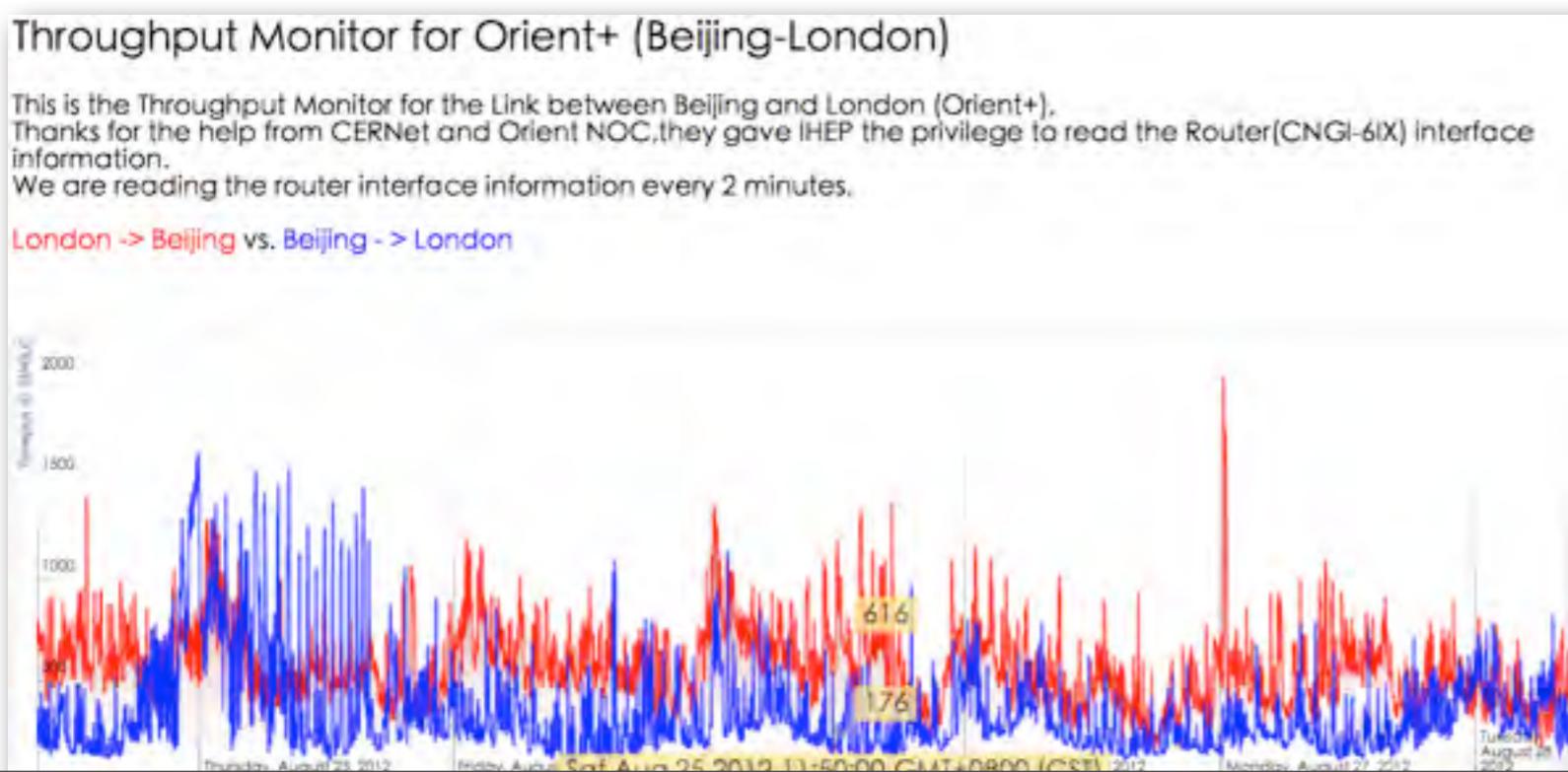
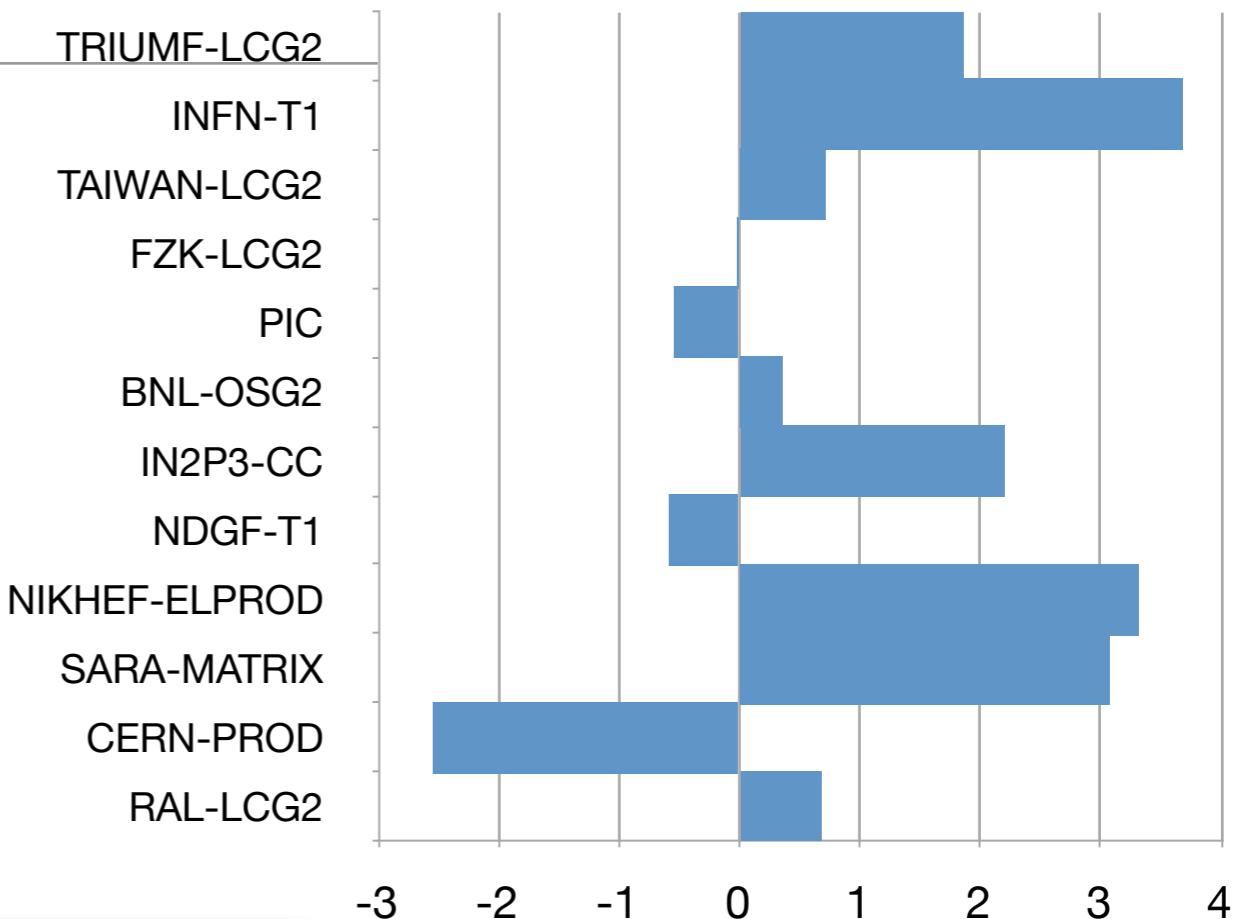
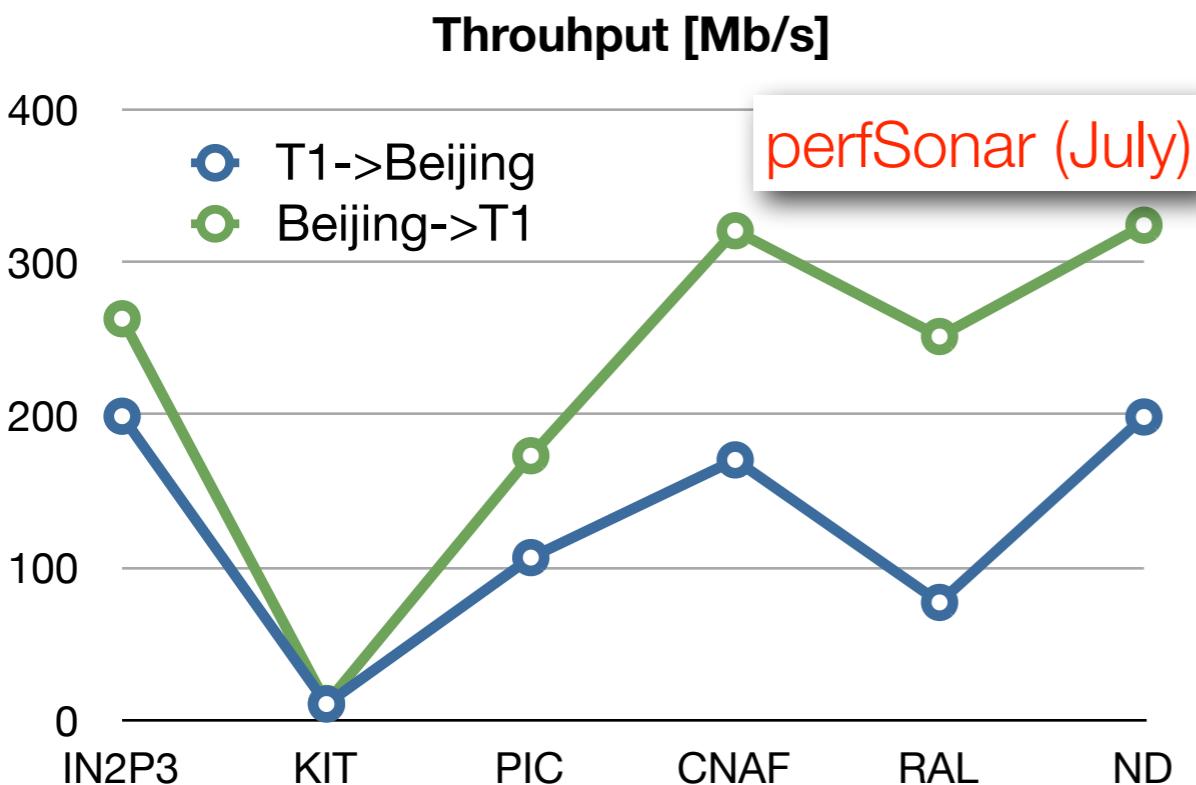
Each T1 is  
different

EU → Tokyo  
better for  
most T1s

- FZK-LCG2 - TOKYO-LCG2 (14361 files)
- IN2P3-CC - TOKYO-LCG2 (83502 files)
- RAL-LCG2 - TOKYO-LCG2 (16632 files)
- CERN-PROD - TOKYO-LCG2 (6937 files)
- SARA-MATRIX - TOKYO-LCG2 (5602 files)
- INFN-T1 - TOKYO-LCG2 (13600 files)
- PIC - TOKYO-LCG2 (7194 files)
- NDGF-T1 - TOKYO-LCG2 (26763 files)

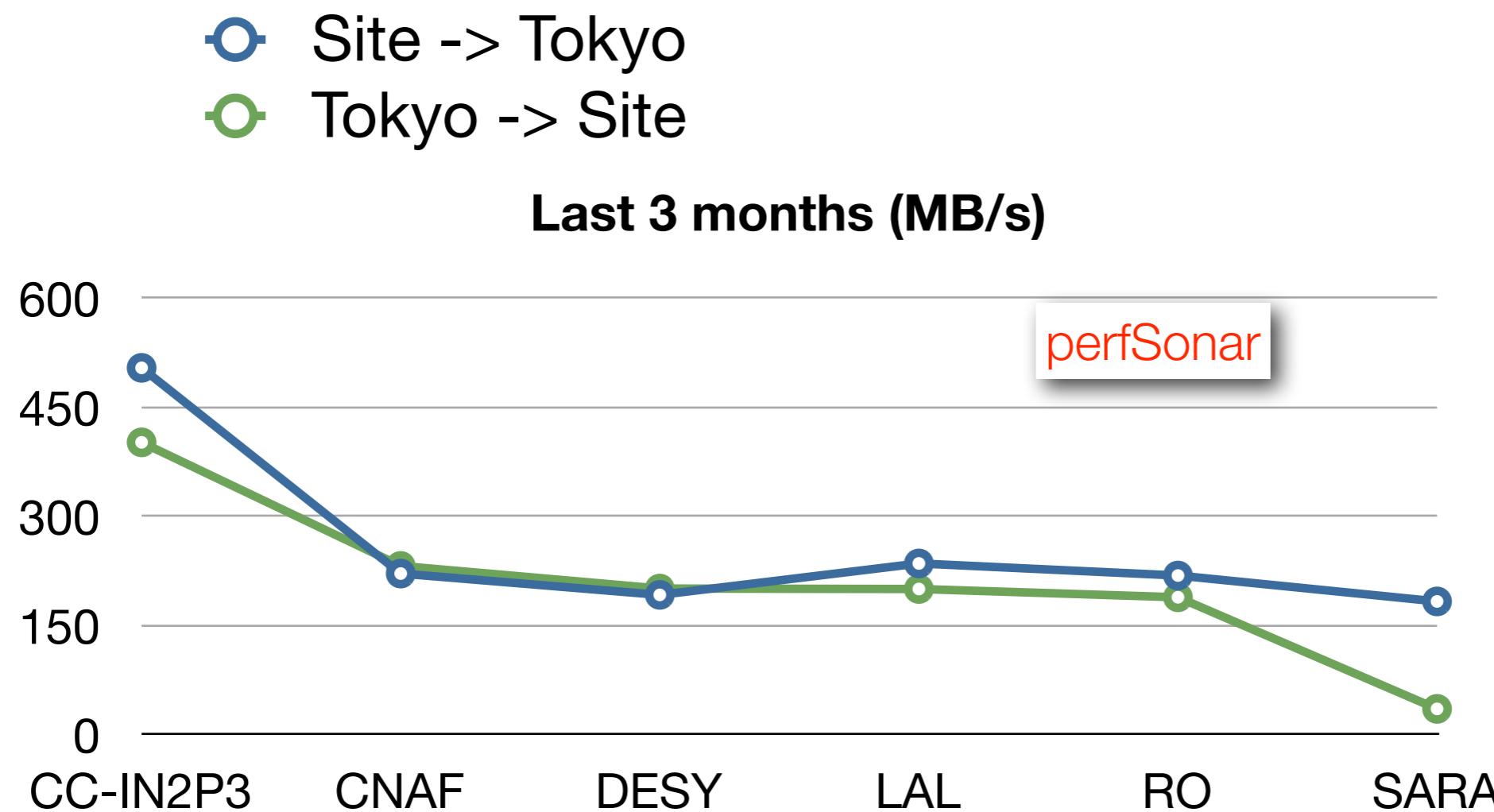


# Beijing - T1s asymmetry

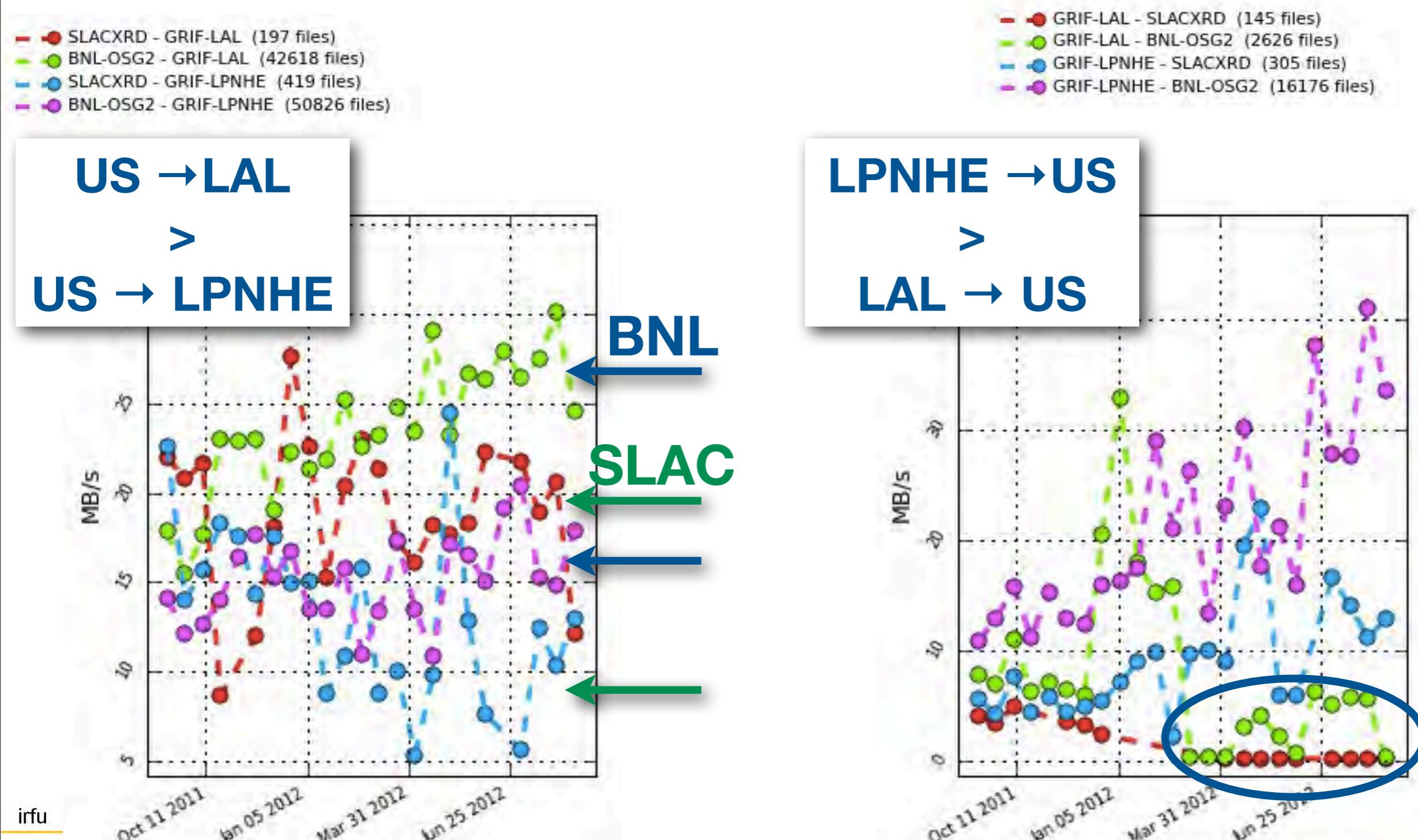


**Beijing -> T1s**  
>  
**T1s -> Beijing**

# Tokyo ↔ EU as seen by perfSonar



# US (LHCONE) ↔ GRIF (LHCONE)



# DISTRIBUTED STORAGE / REMOTE ACCESS

---

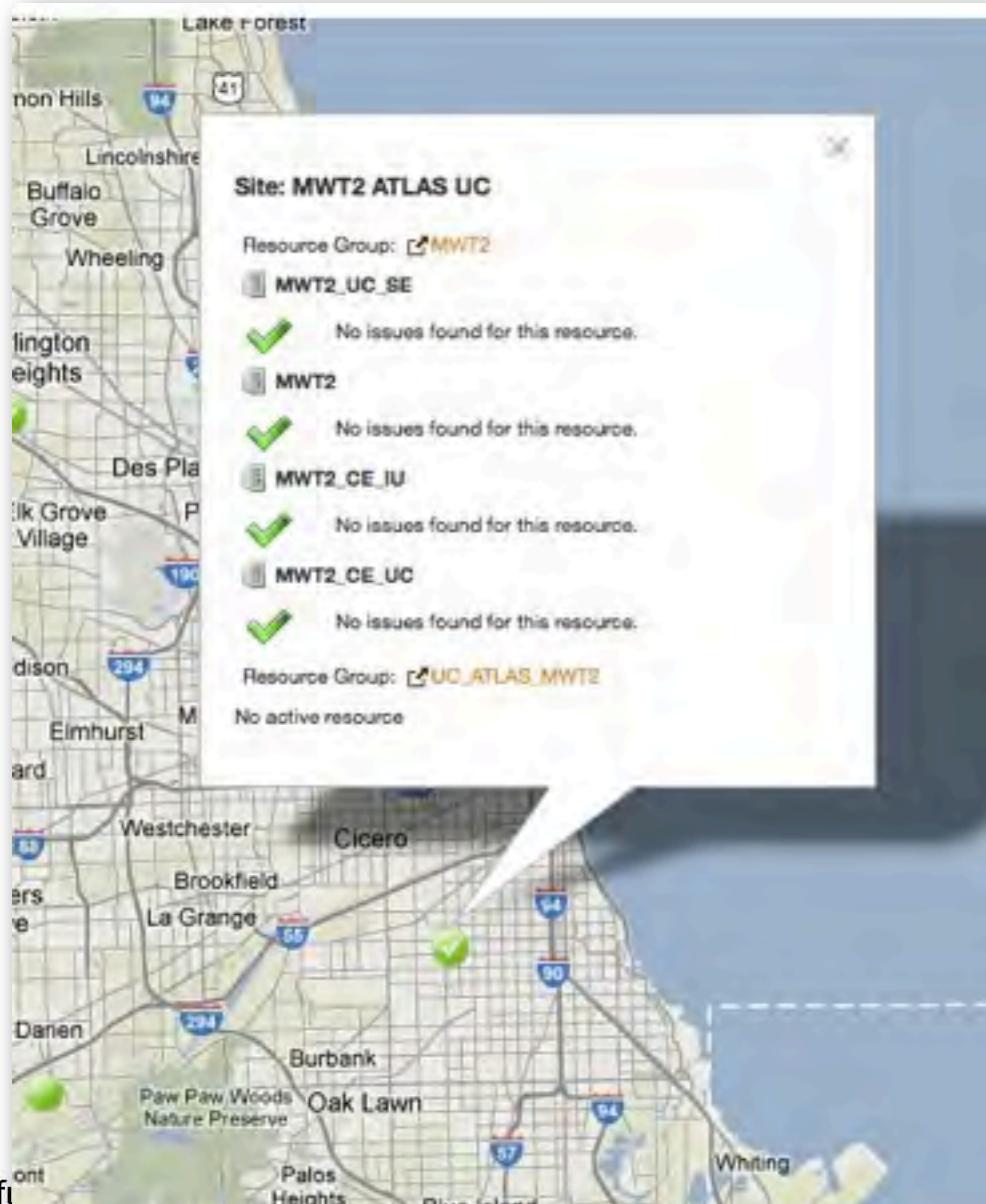
- Better used of storage resources (disk prices!)
- Simplification of data management
- Eventually remote access (with caching at both ends); direct reading or file copy
- Bandwidth and stability needed

On going projects

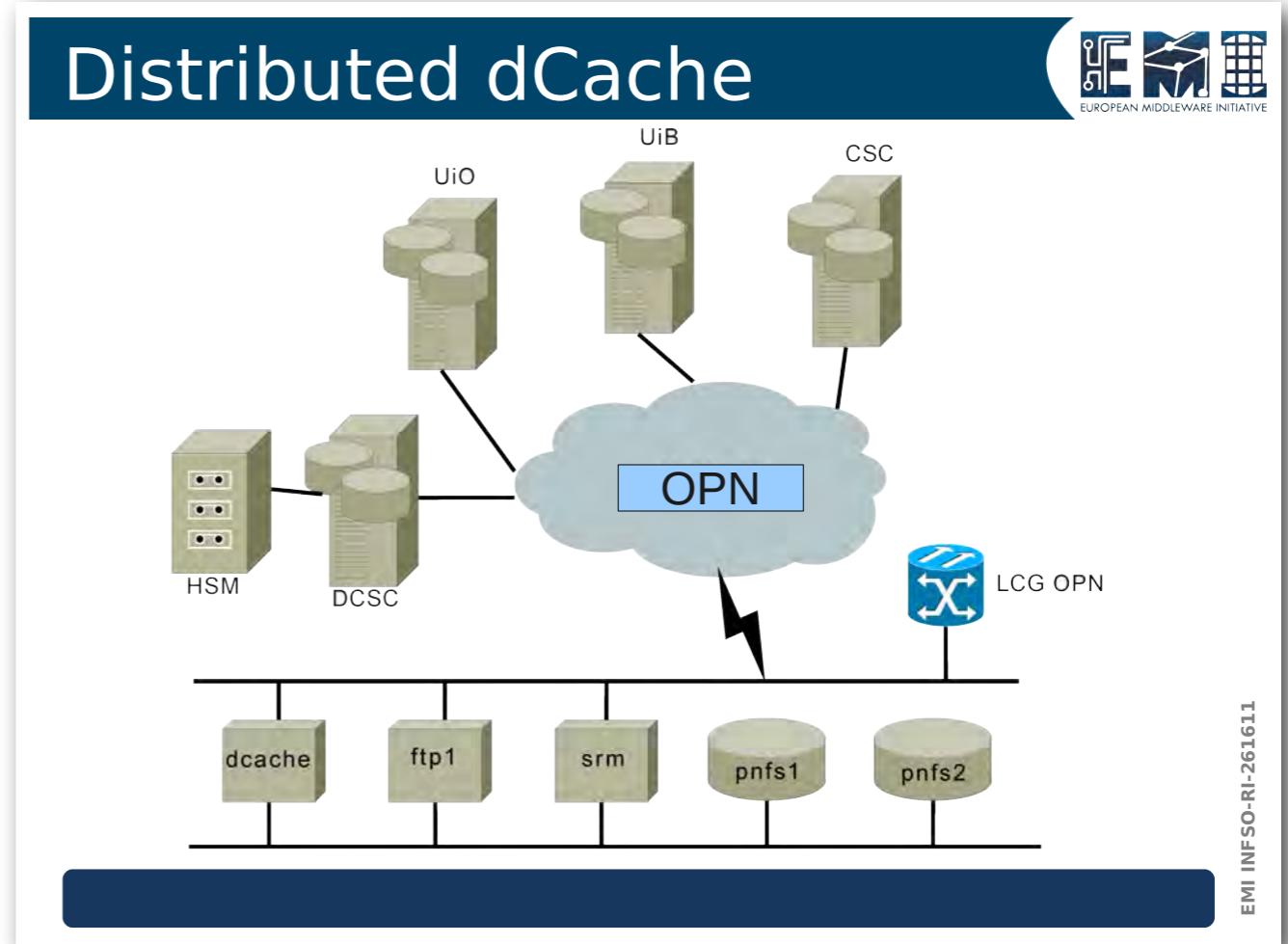
- Storage : dCache, dpm,...
- Protocol : Xrootd, HTTP/WebDAV

# Existing distributed dCache systems : 2 examples

## MWT2 (Chicago)



## NORDUGRID

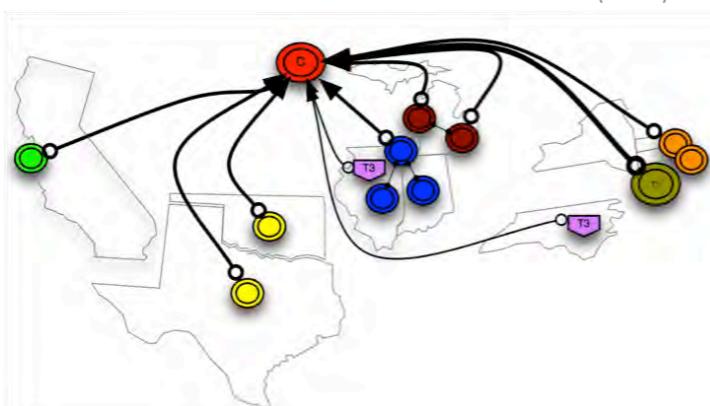


# Xrootd federation project in US

REMOTE ACCESS

## R&D Activity to Production

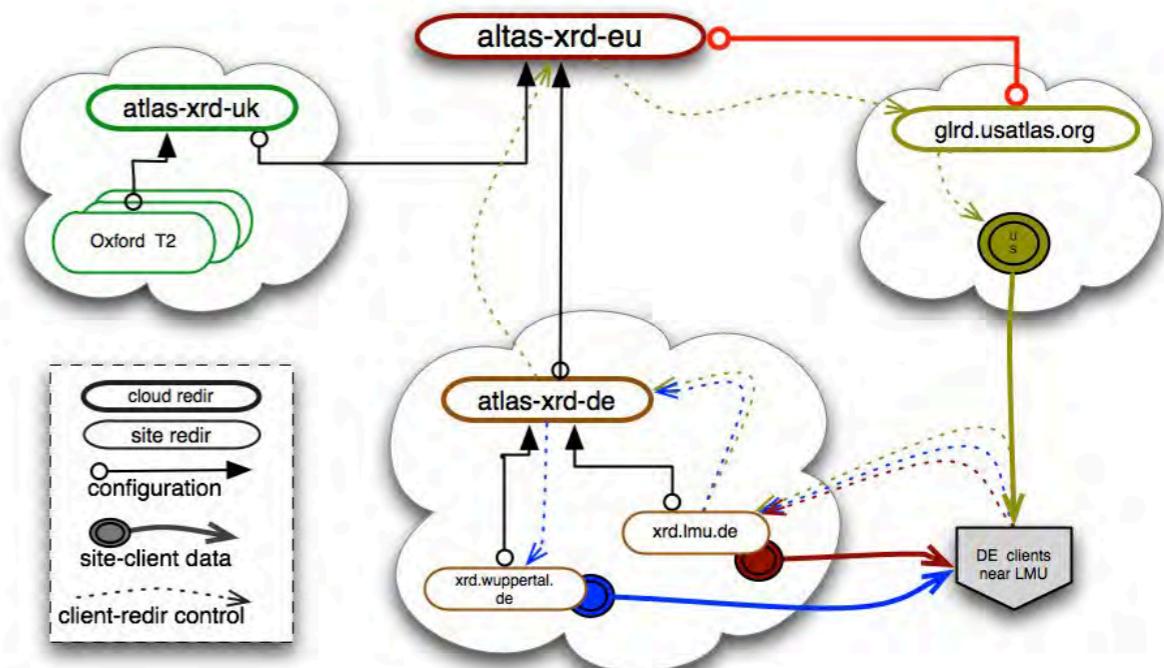
- 2011 R&D project FAX (Federating ATLAS data stores using Xrootd) was deployed over US Tier 1, Tier 2s and some Tier3s
- Feasibility testing monitoring, site integrations
- In June 2012 extended effort to European sites as an ATLAS-wide project



2

## EU federation tests

Four levels of redirection:  
site-cloud-zone-global

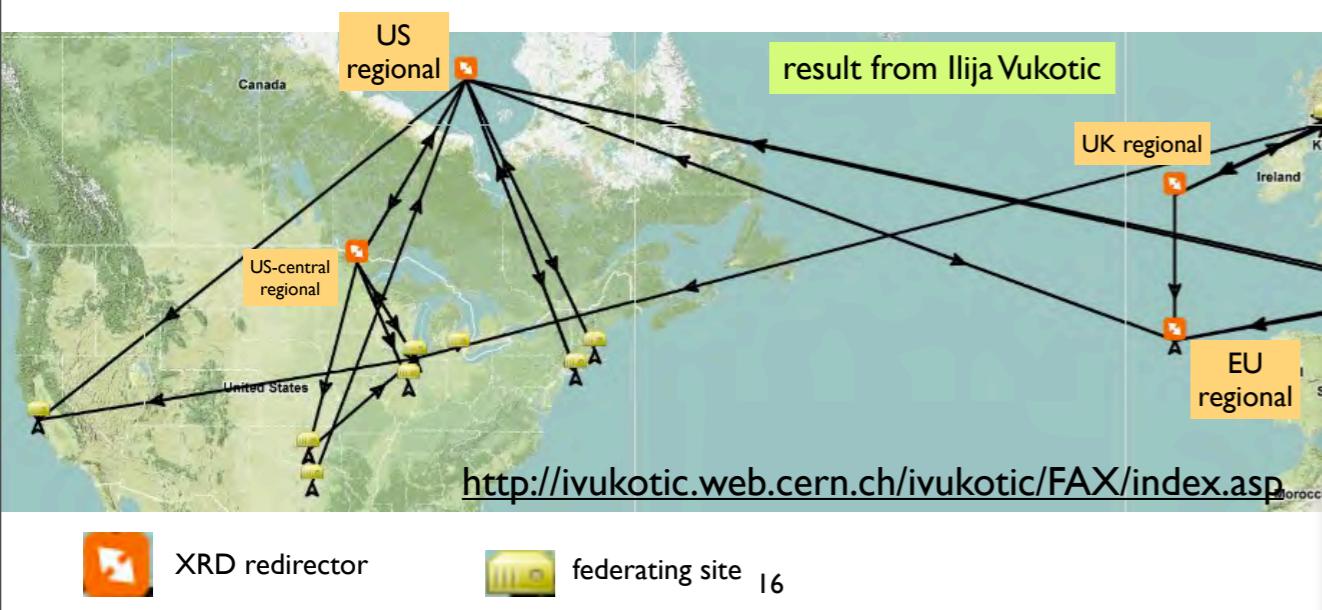


Start locally - expand search as needed

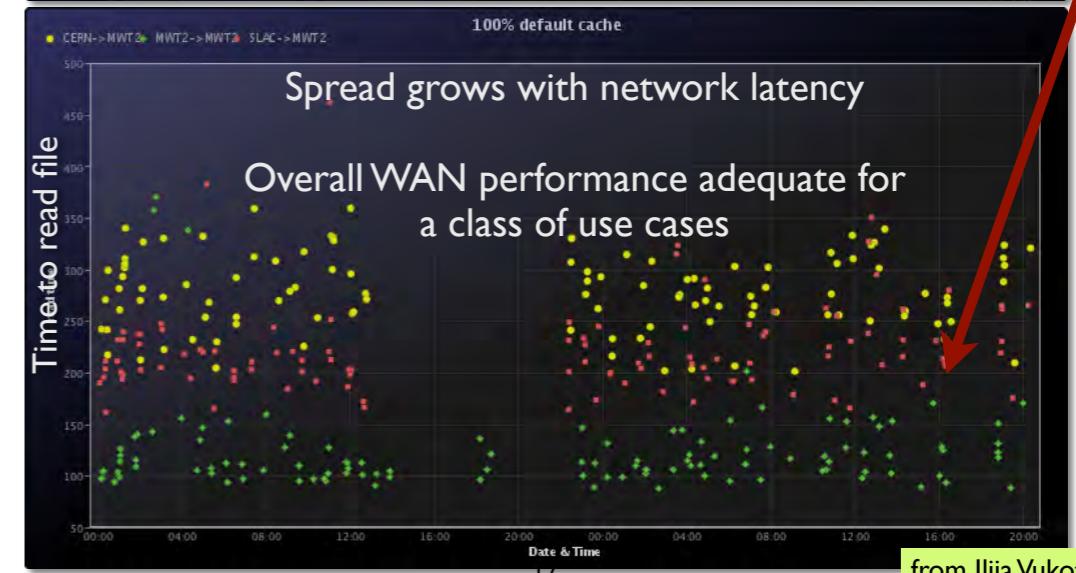
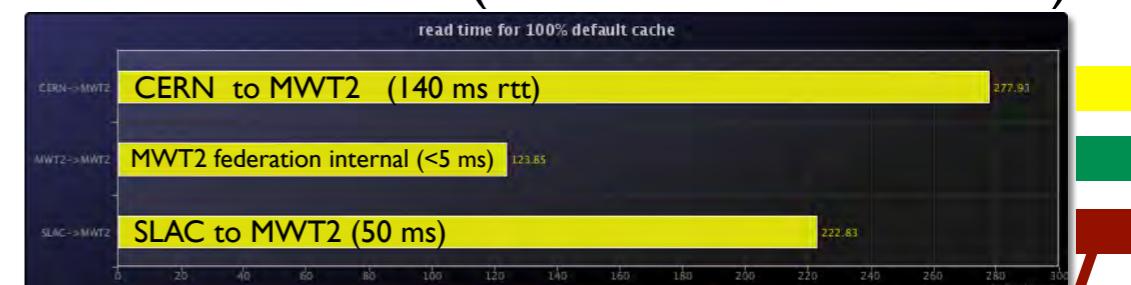
15

# Topology validation

- Launch jobs to every site, test reading of site-specific files at every other site
- Parse client logs to infer resulting redirection



## WAN Read Tests (basis for “cost matrix”)



## HTTP/WebDAV

Storage Federations using standard web protocols

dCache.org

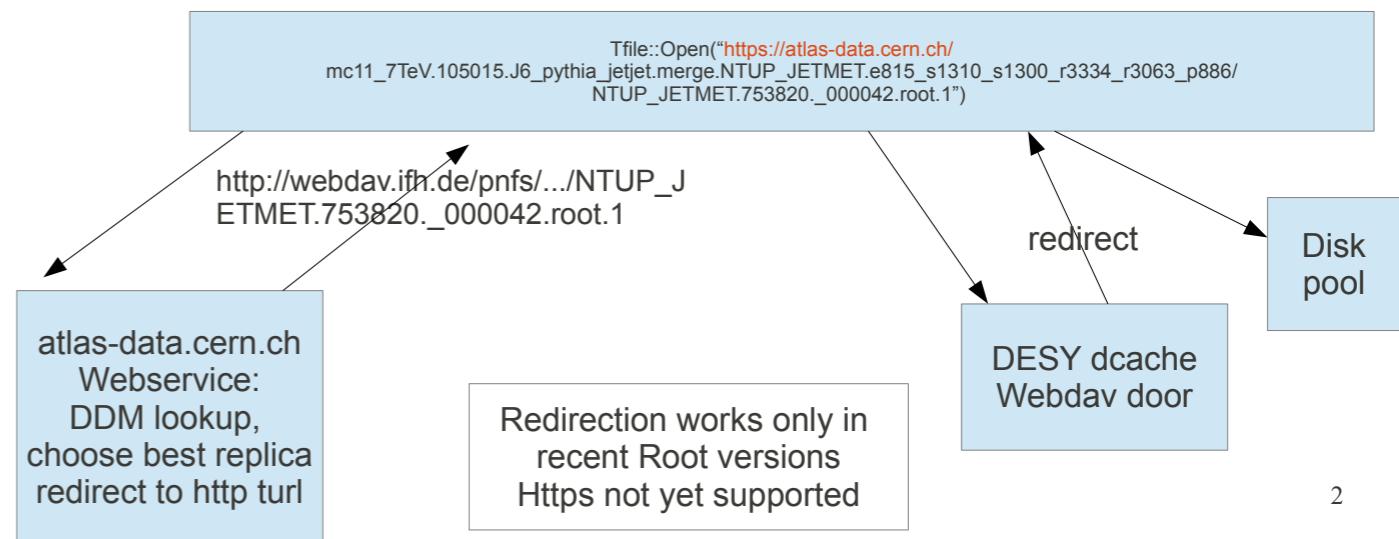


- Project with CERN DM under the umbrella of EMI but not limited to the EMI funding period.
- Definition of TEG:
  - “Collection of disparate storage resources managed by co-operating but independent administrative domains transparently accessible via a common name space”
- We do it with standard HTTP/WebDAV

Atlas WS 2012, CERN| dCache.org| 10 Sep 2012 | 11

### Use http urls for input files

- DDM enabled web redirection service
  - generic url including dataset and lfn
  - redirects to http turl in dcache/dpm storage



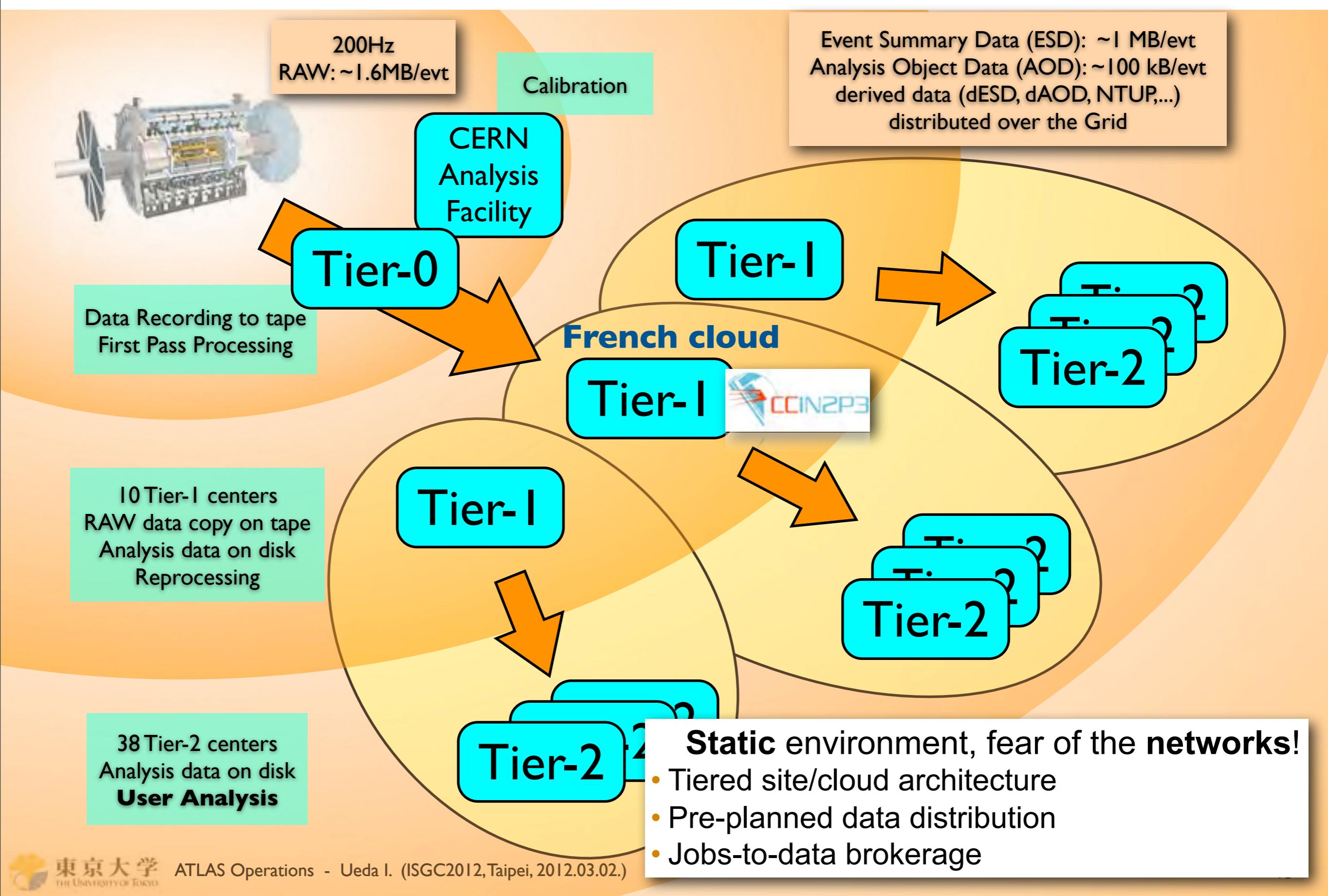
# Summary

---

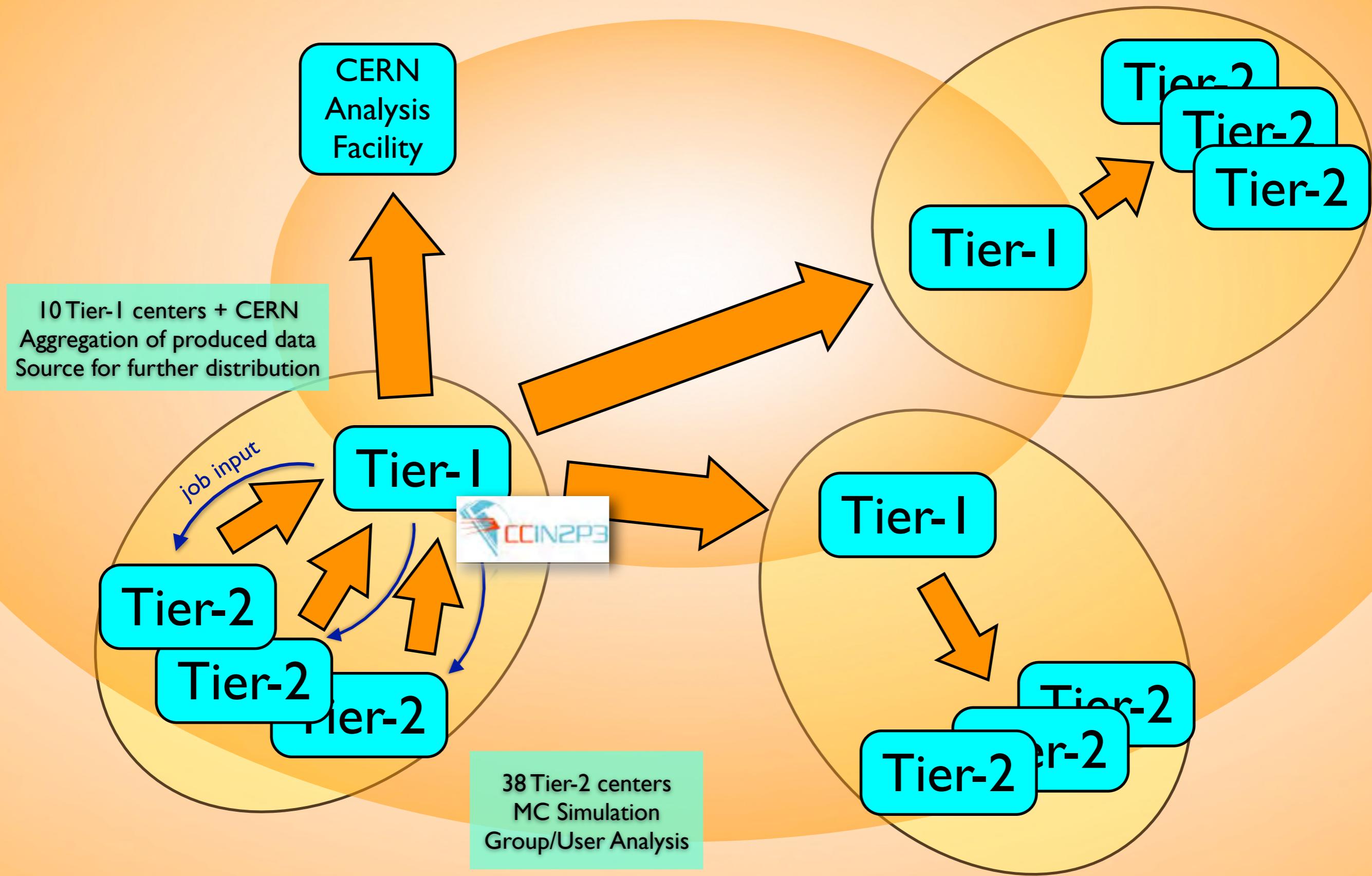
- Thanks to the high performances of networks
- ATLAS computing model has changed significantly: simplification of data and workflow management
- Would have been impossible to handle current data volume (LHC performing beyond expectations) and LHC running extension up to spring 2013 with initial model
- More efficient use of storage resources (reduce replica counts; direct sharing of replicas across sites)
- Ongoing projects (distributed storage, remote access) will further change the landscape

# BACKUP

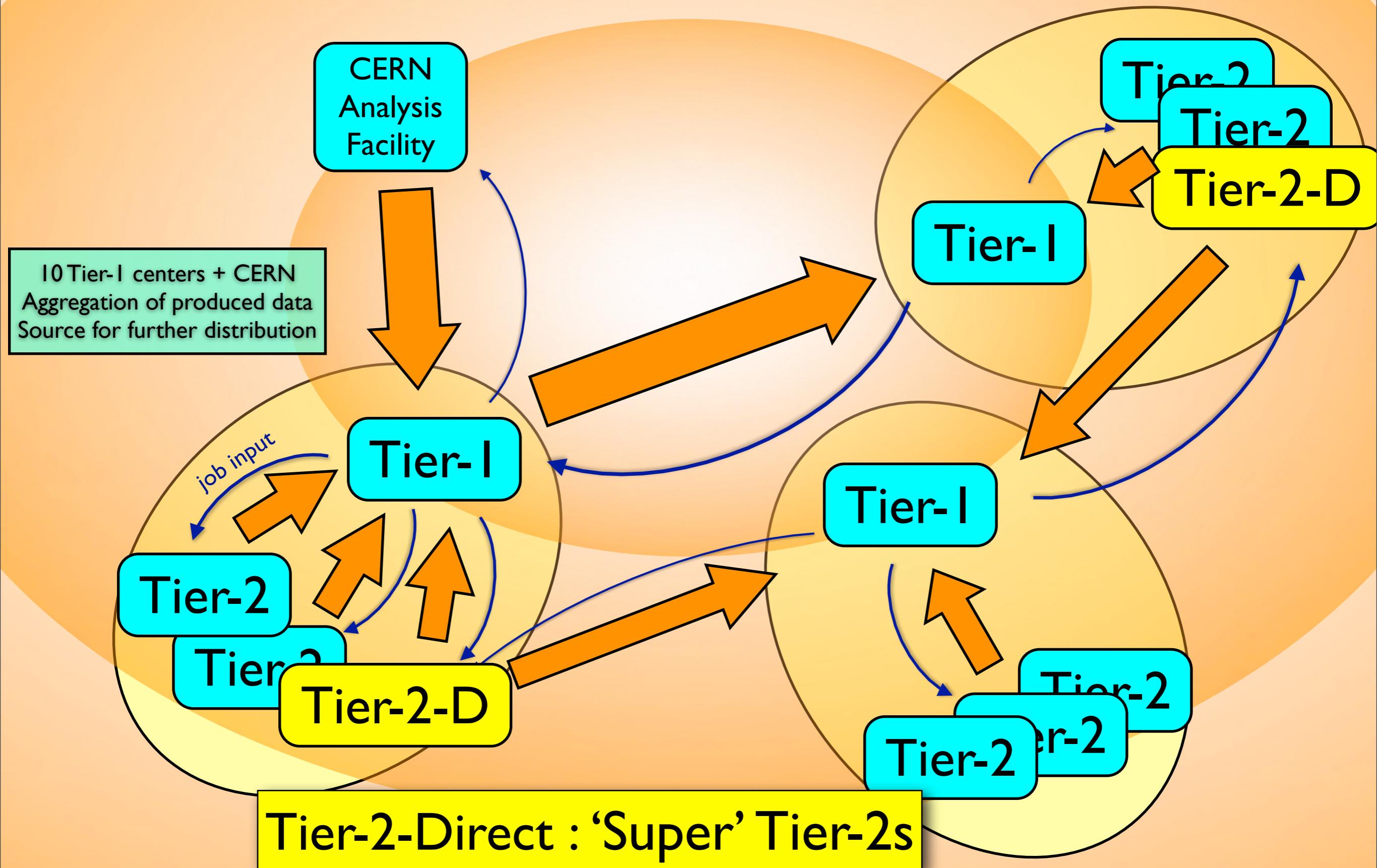
# ATLAS Computing Model: T0 Data Flow



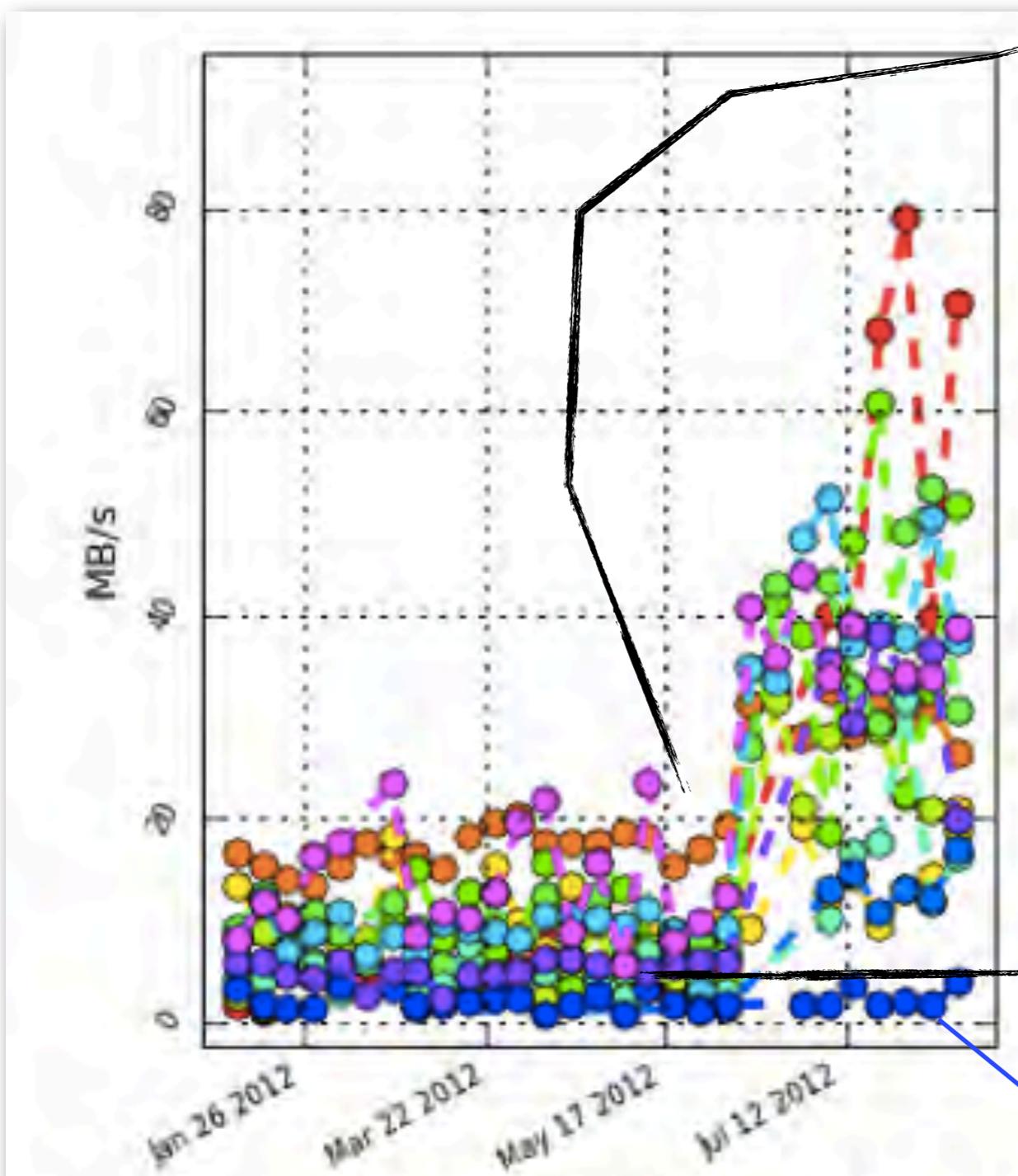
# ATLAS Computing Model: MC Data Flow



# Data Processing Model Revised



# T1s -> IN2P3-CPPM



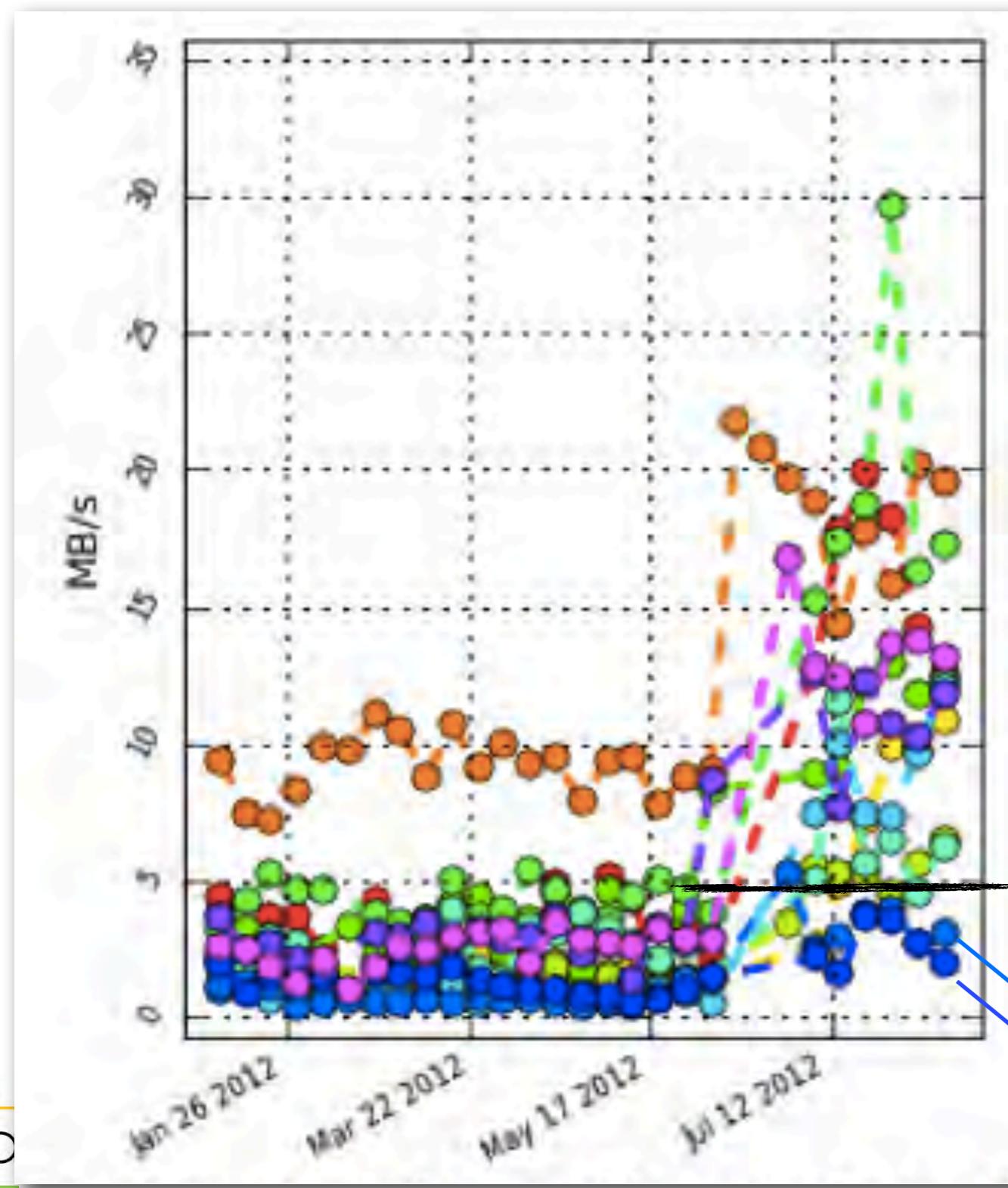
June 25th  
connected to LHCONe  
@ 10 Gb/s

- FZK-LCG2 - IN2P3-CPPM (1360 files)
- IN2P3-CC - IN2P3-CPPM (106227 files)
- RAL-LCG2 - IN2P3-CPPM (2017 files)
- BNL-OSG2 - IN2P3-CPPM (4383 files)
- SARA-MATRIX - IN2P3-CPPM (2031 files)
- INFN-T1 - IN2P3-CPPM (1898 files)
- PIC - IN2P3-CPPM (541 files)
- NDGF-T1 - IN2P3-CPPM (1571 files)
- TAIWAN-LCG2 - IN2P3-CPPM (275 files)
- TRIUMF-LCG2 - IN2P3-CPPM (186 files)
- NIKHEF-ELPROD - IN2P3-CPPM (292 files)
- CERN-PROD - IN2P3-CPPM (3122 files)

5 MB/s

TRIUMF

# IN2P3-CPPM → T1s

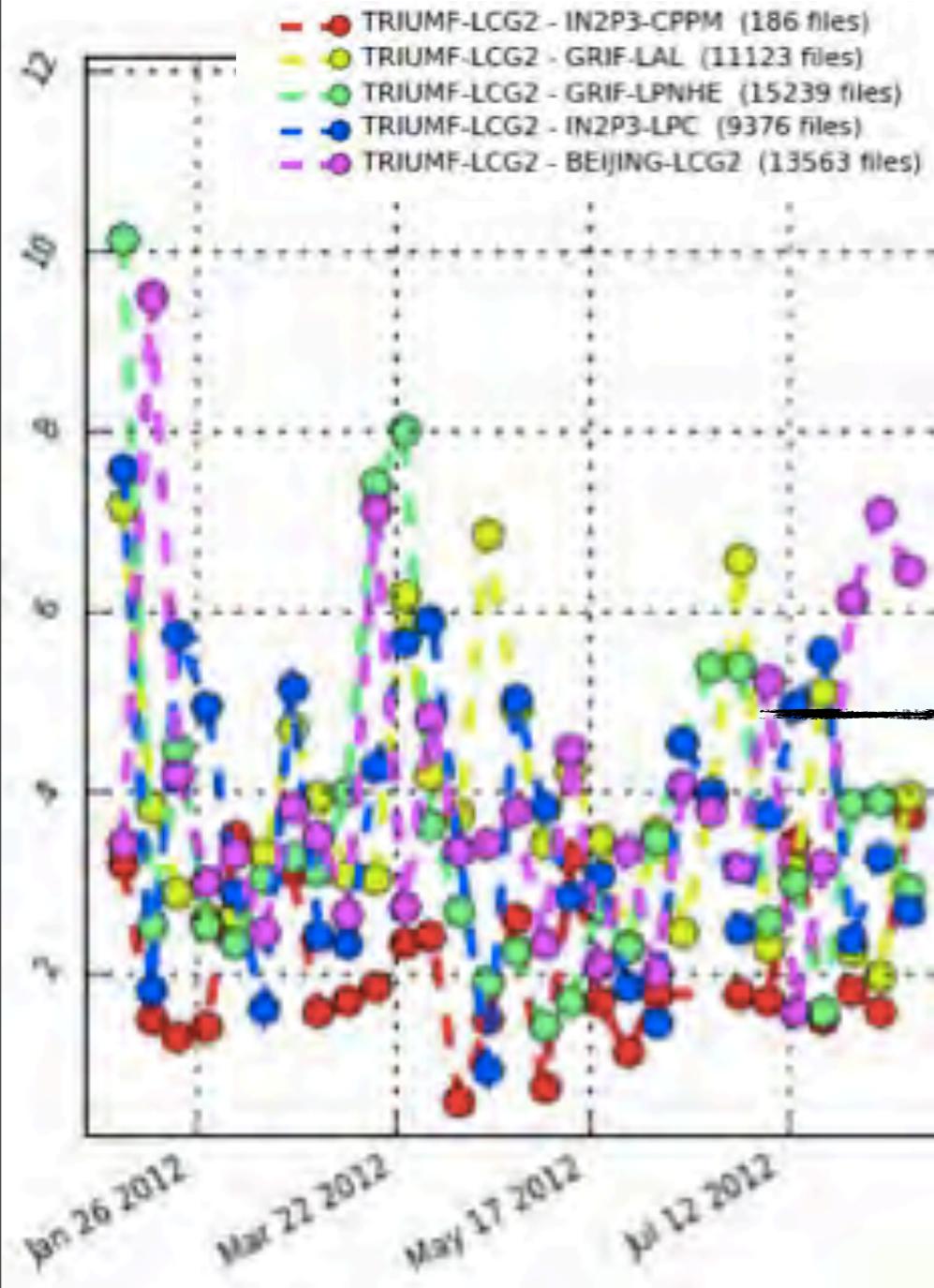


5 MB/s

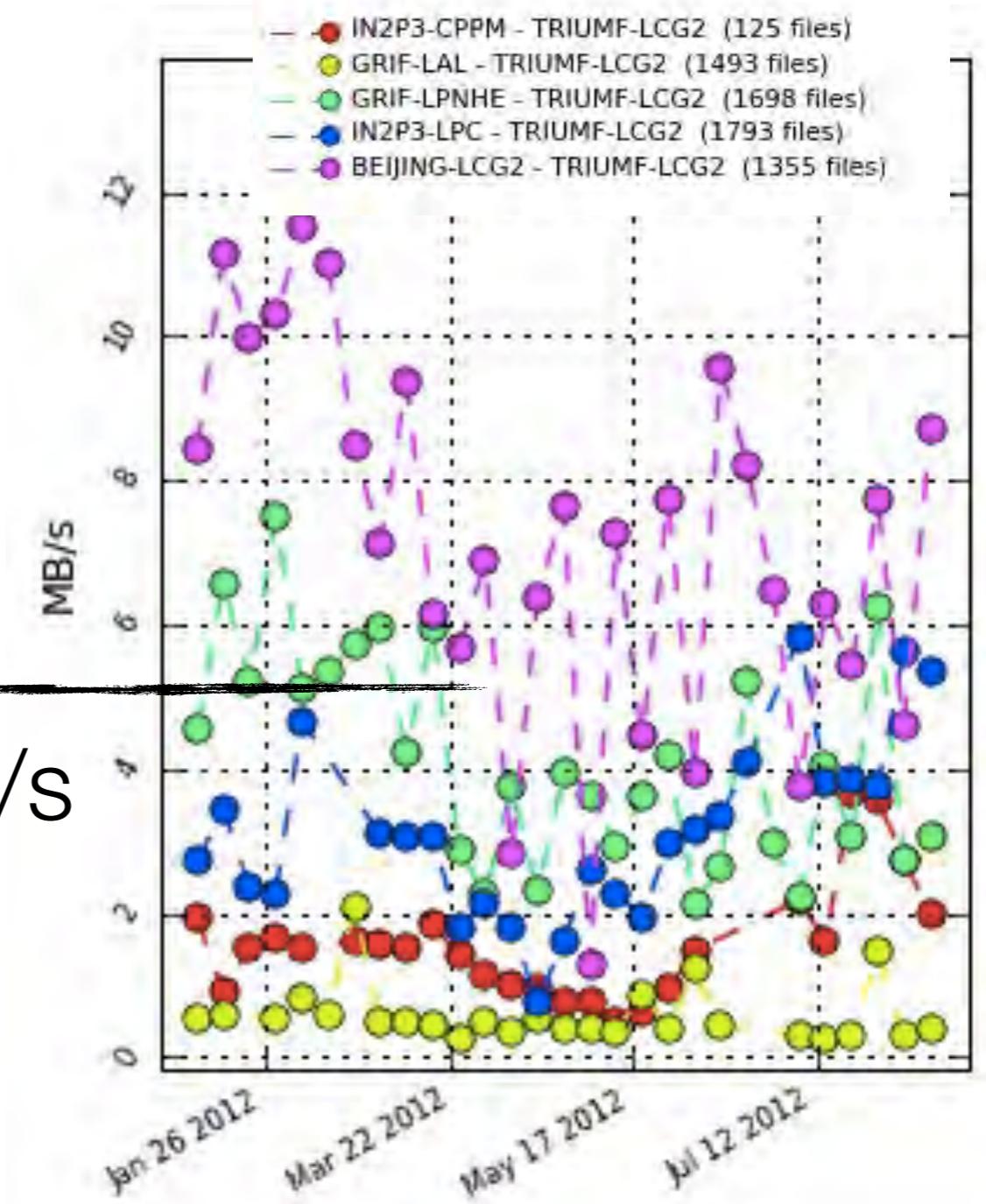
TW  
TRIUMF

# TRIUMF <-> FR T2Ds

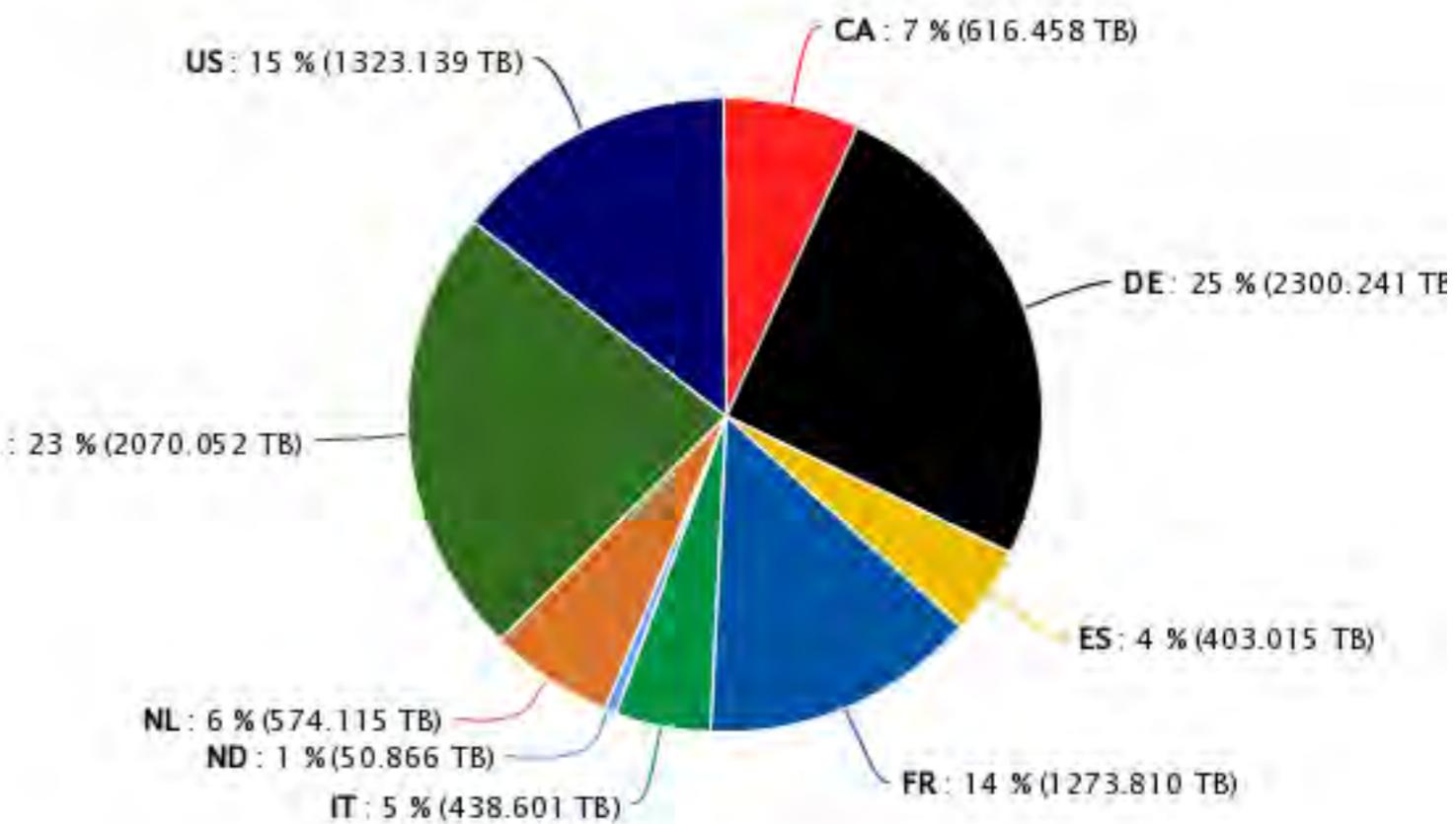
~None of T2Ds ever reaches the 5 MB/s canonical parameter value



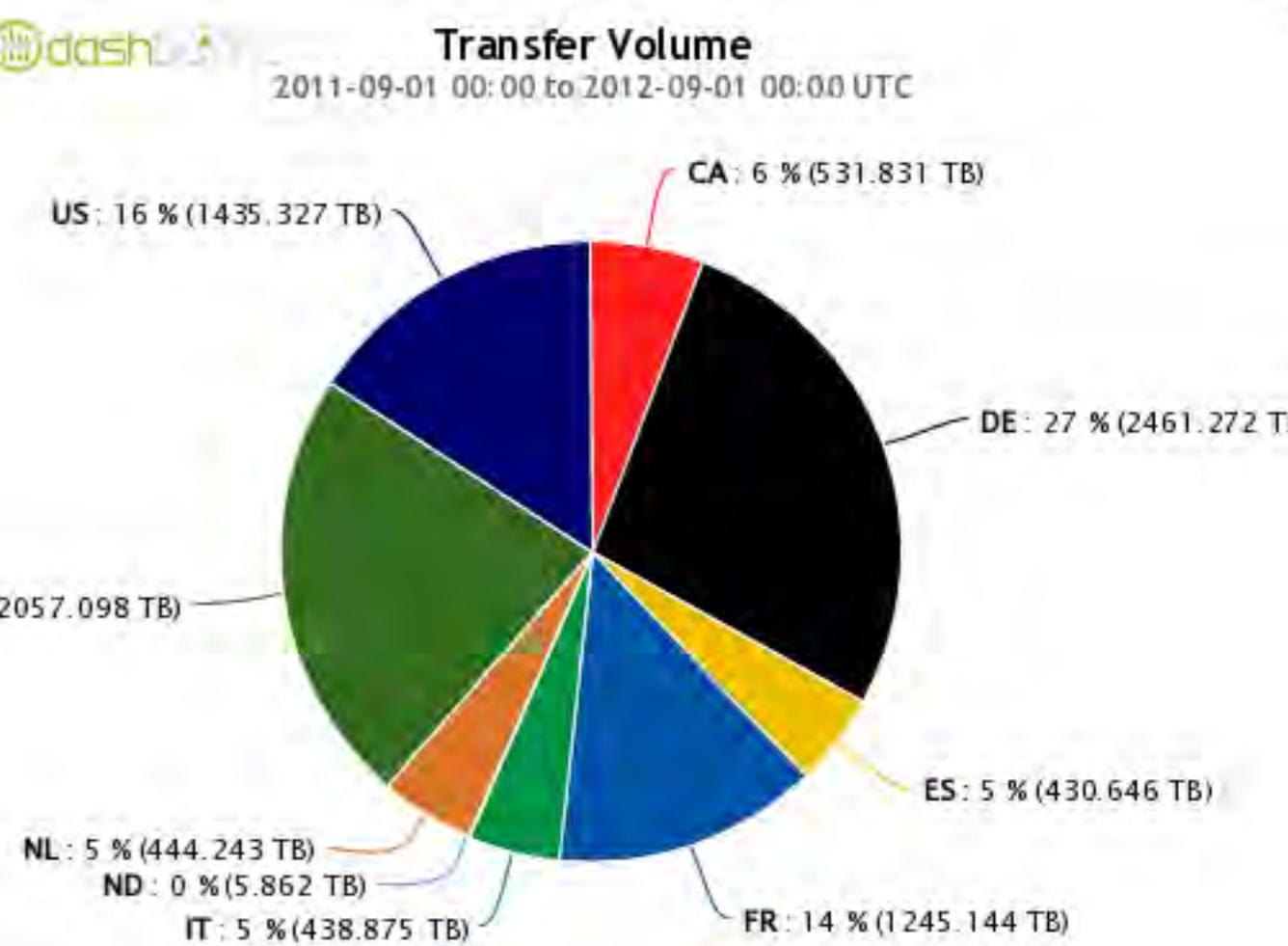
5 MB/s



2011-09-01 00:00 to 2012-09-01 00:00 UTC



## T2-T2 destination

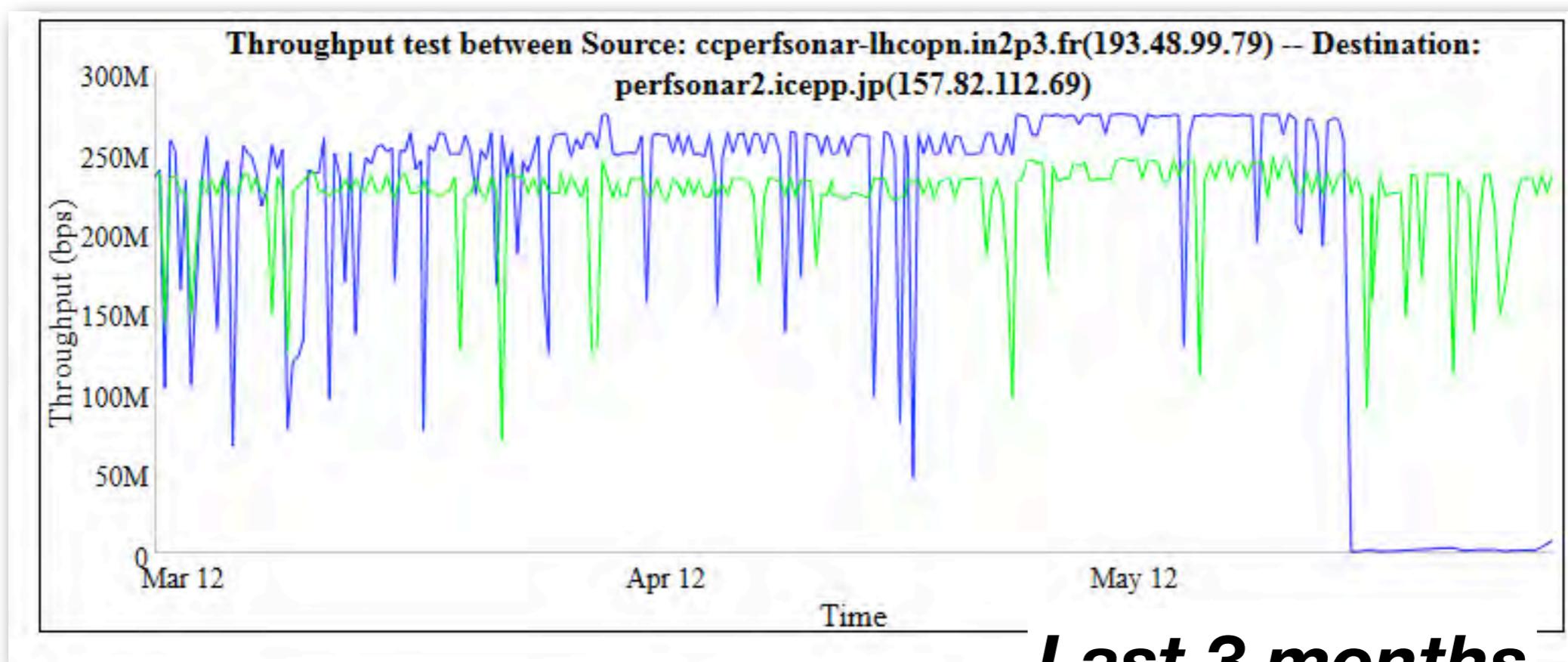


## T2-T2 source

# Network throughput measured with perfSONAR

**CCIN2P3 → Tokyo**

**Tokyo → CCIN2P3**



No so stable  
better by ~5% for CCIN2P3 → Tokyo

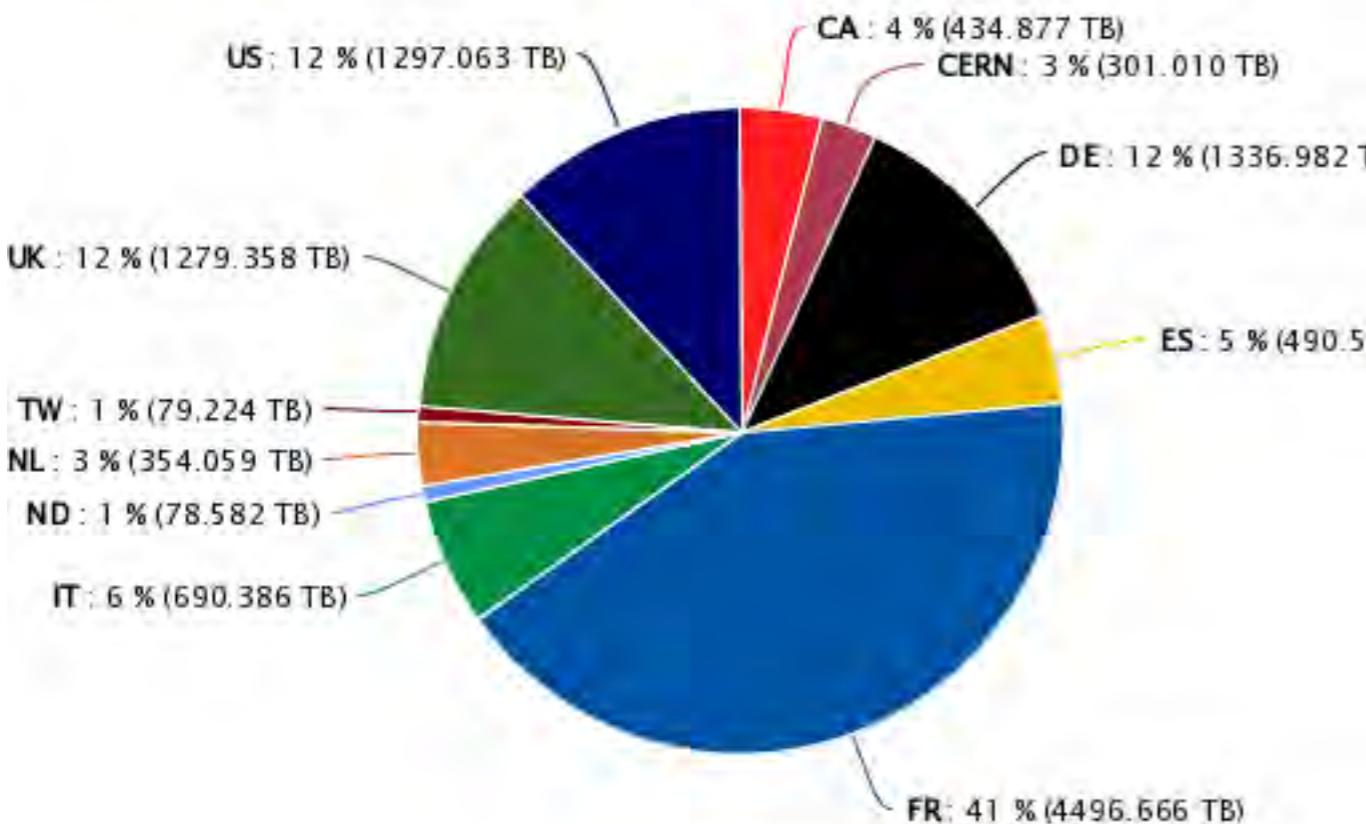


# CCIN2P3 Exports



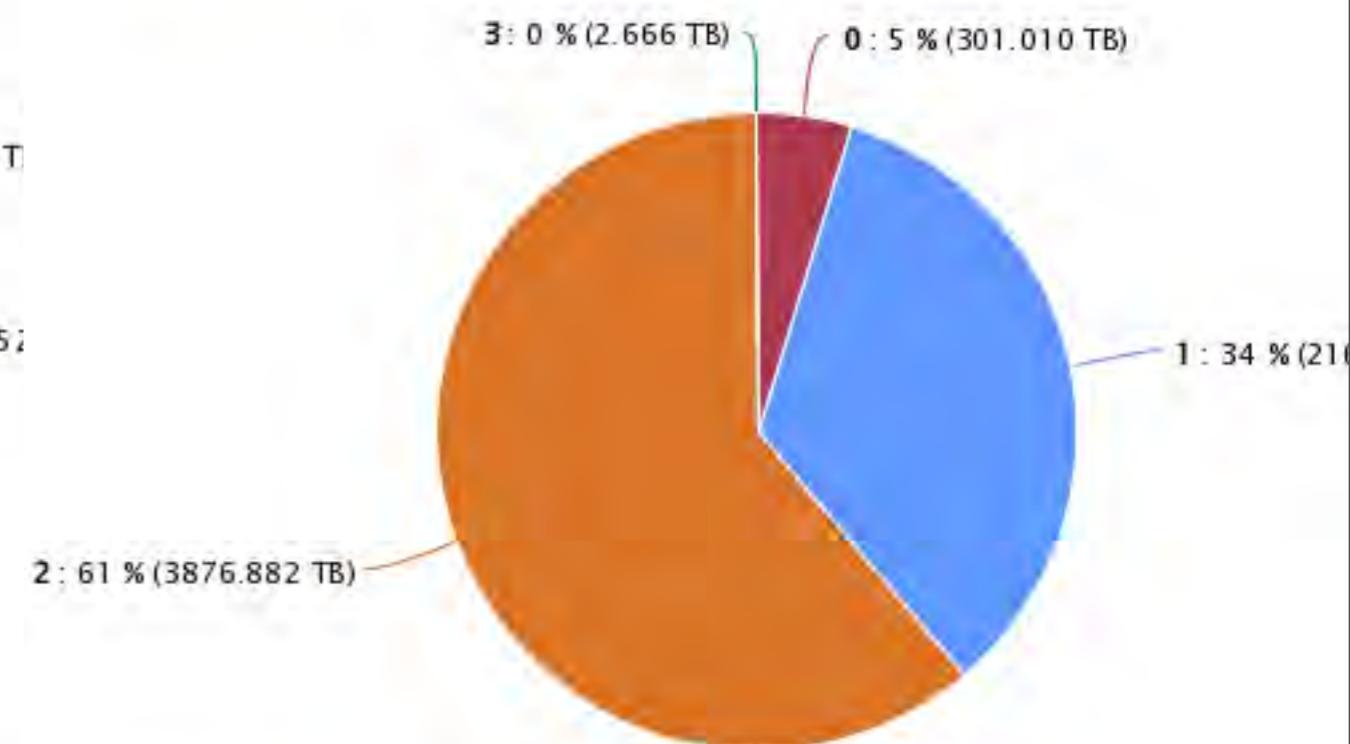
Transfer Volume

2011-09-01 00:00 to 2012-09-01 00:00 UTC



Transfer Volume

2011-09-01 00:00 to 2012-09-01 00:00 UTC

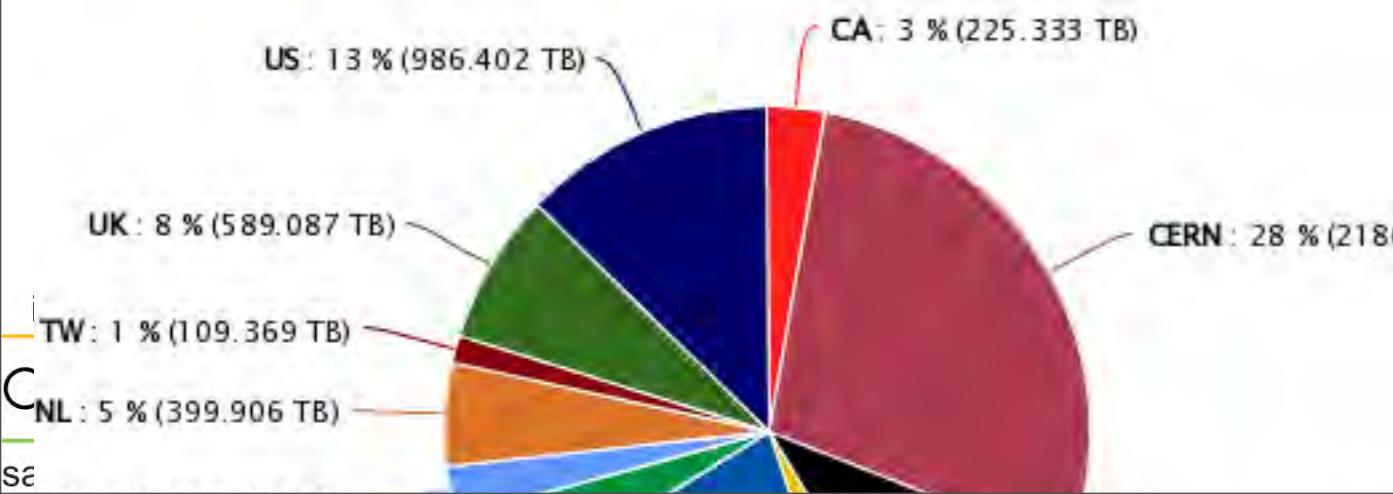


To Lyon



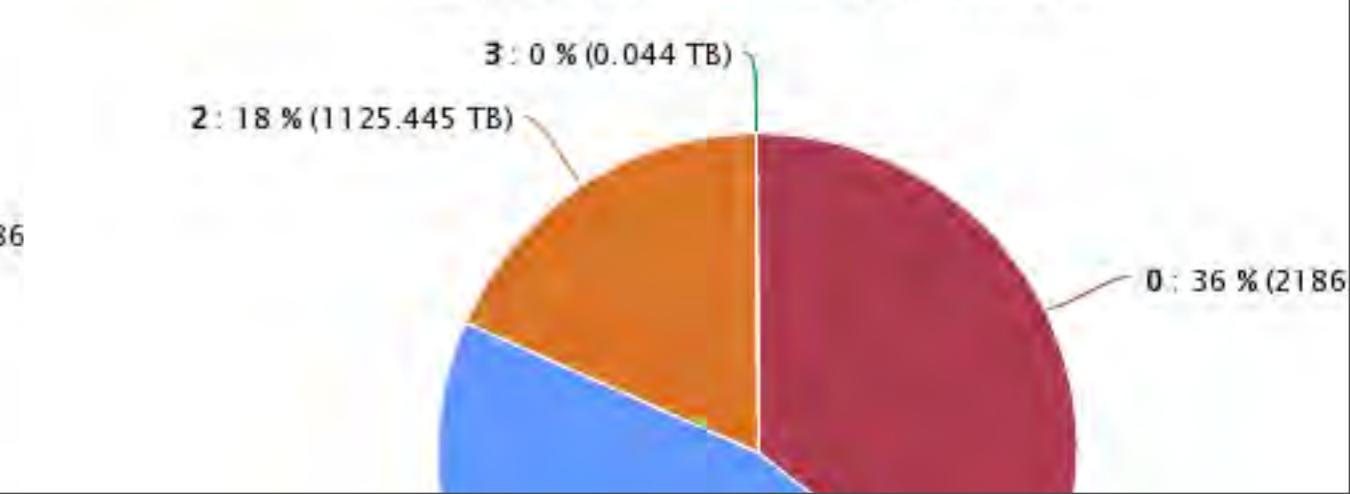
Transfer Volume

2011-09-01 00:00 to 2012-09-01 00:00 UTC



Transfer Volume

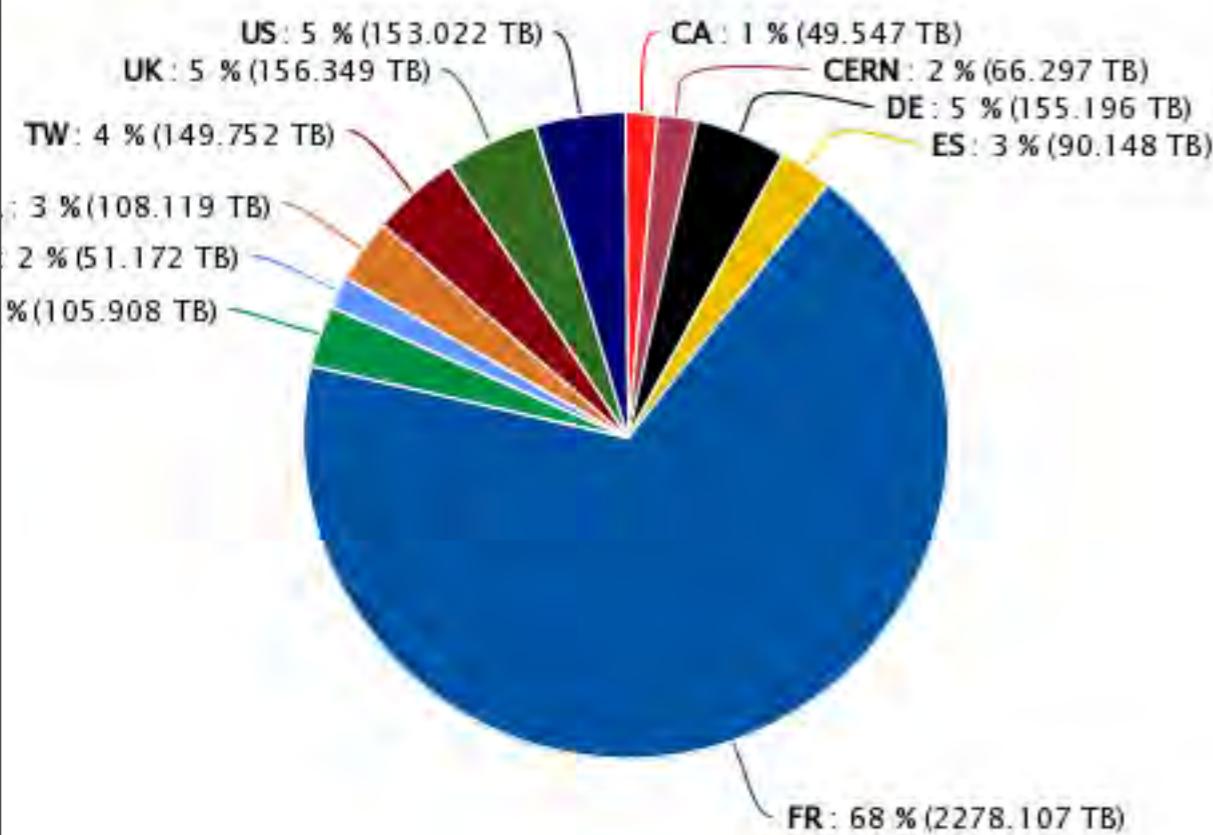
2011-09-01 00:00 to 2012-09-01 00:00 UTC



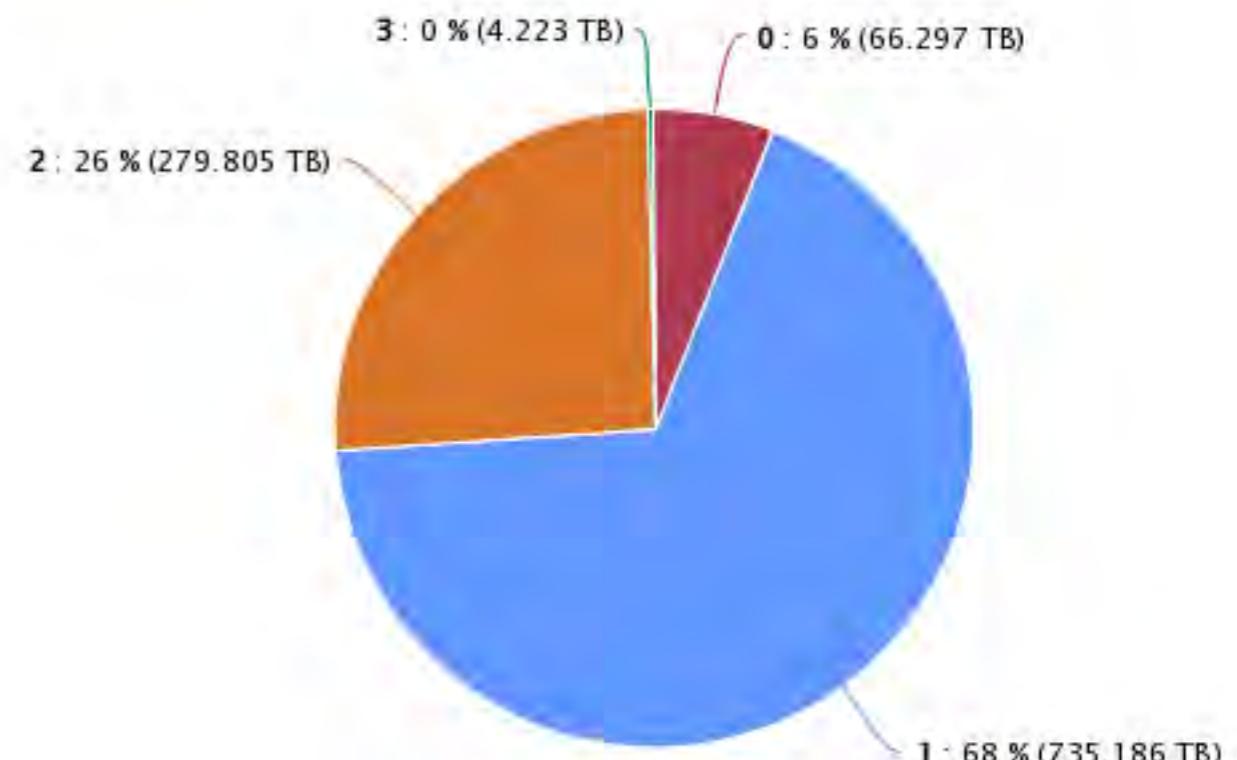
# T2s Exports



**Transfer Volume**  
2011-09-01 00:00 to 2012-09-01 00:00 UTC



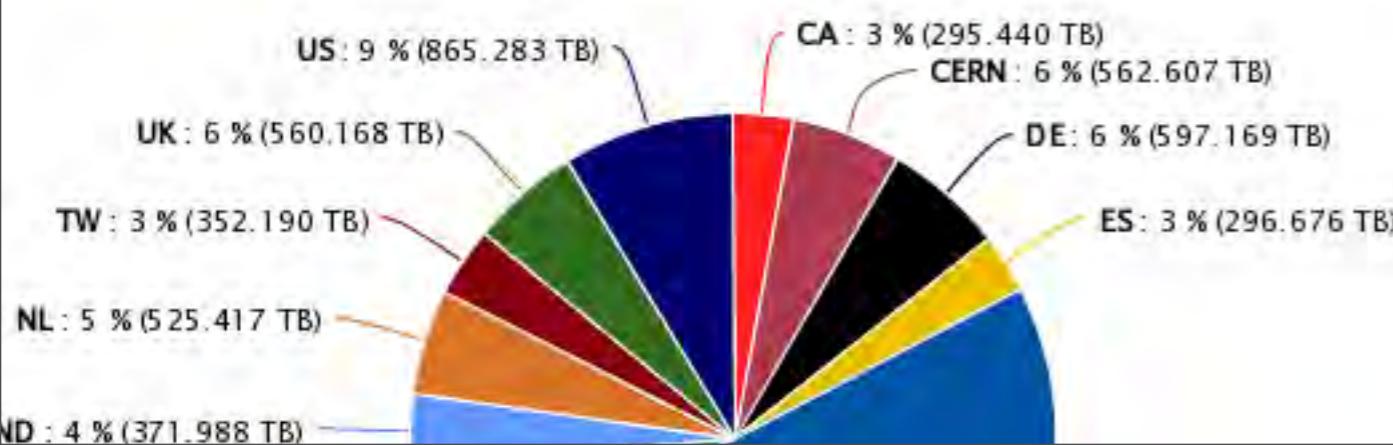
**Transfer Volume**  
2011-09-01 00:00 to 2012-09-01 00:00 UTC



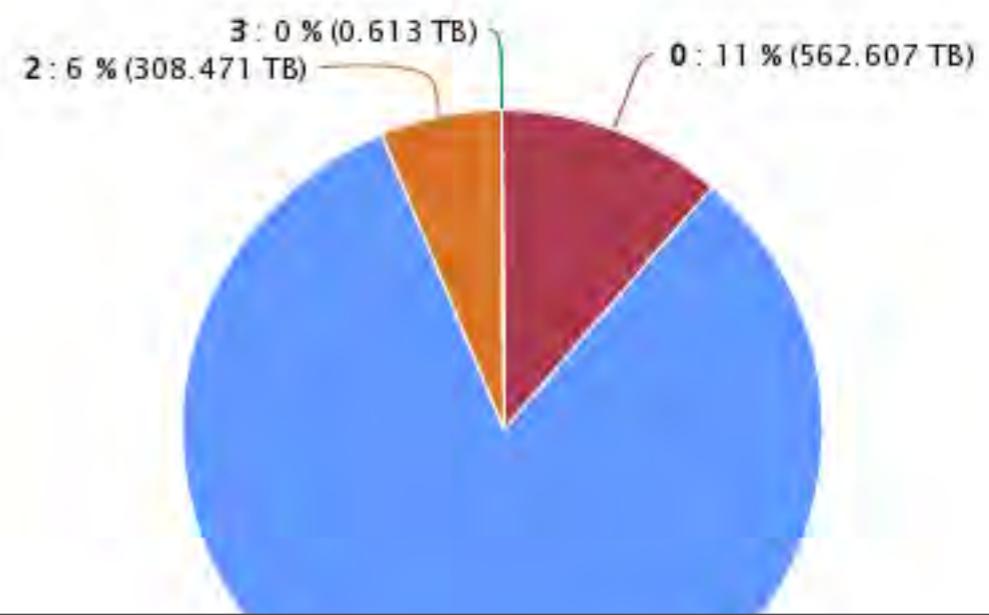
# T2s Import



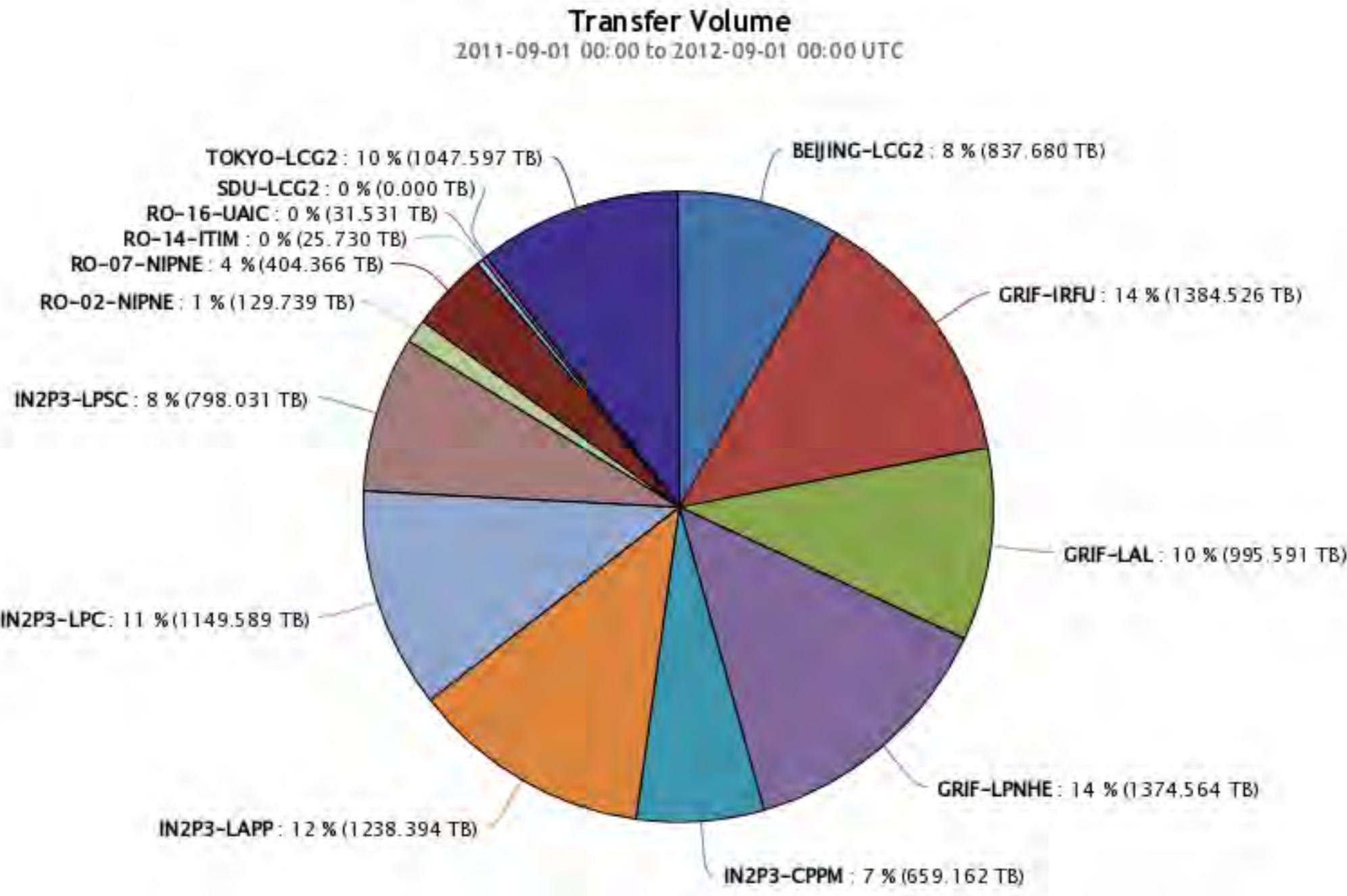
**Transfer Volume**  
2011-09-01 00:00 to 2012-09-01 00:00 UTC



**Transfer Volume**  
2011-09-01 00:00 to 2012-09-01 00:00 UTC

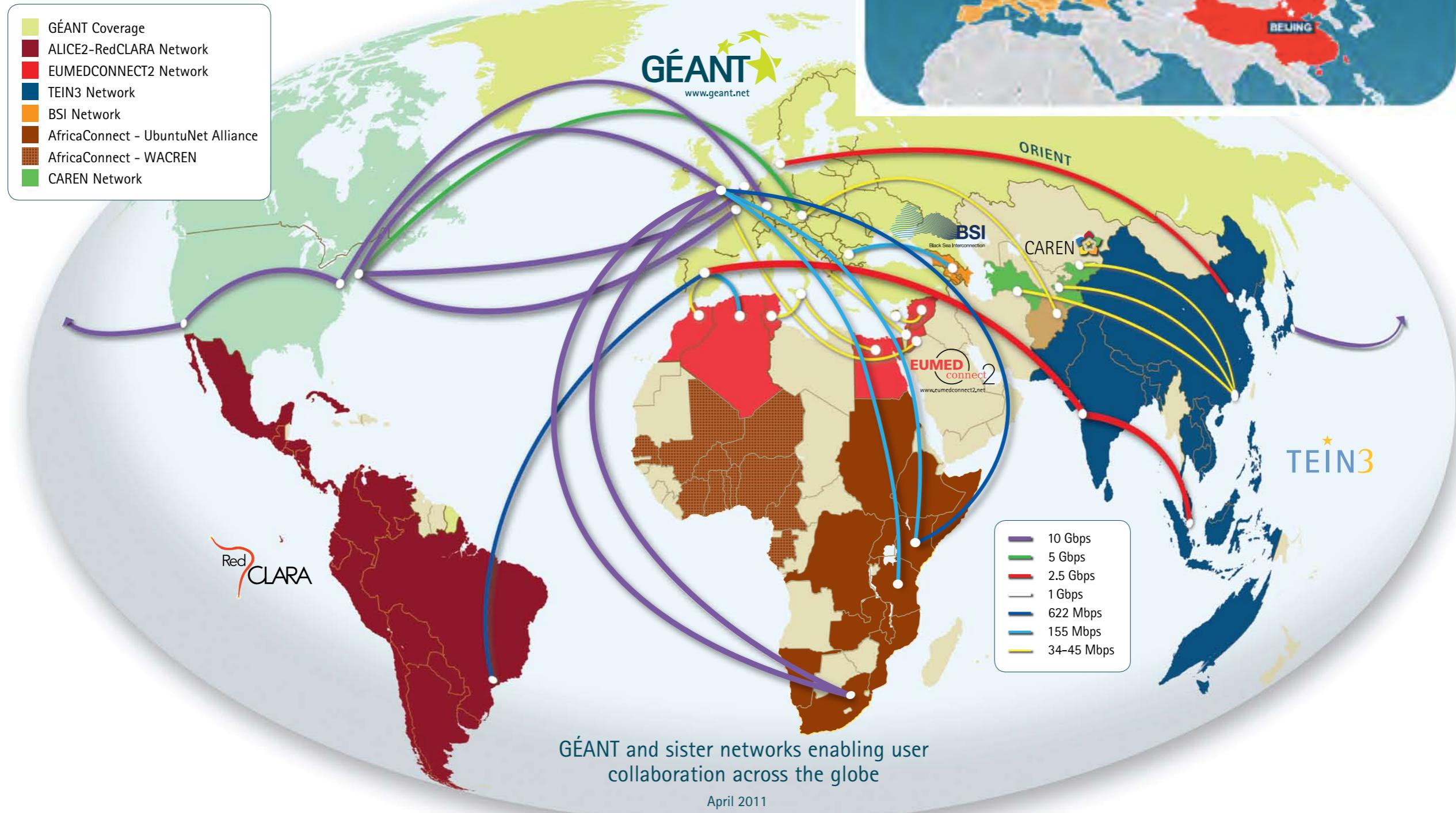


# destination on FR cloud



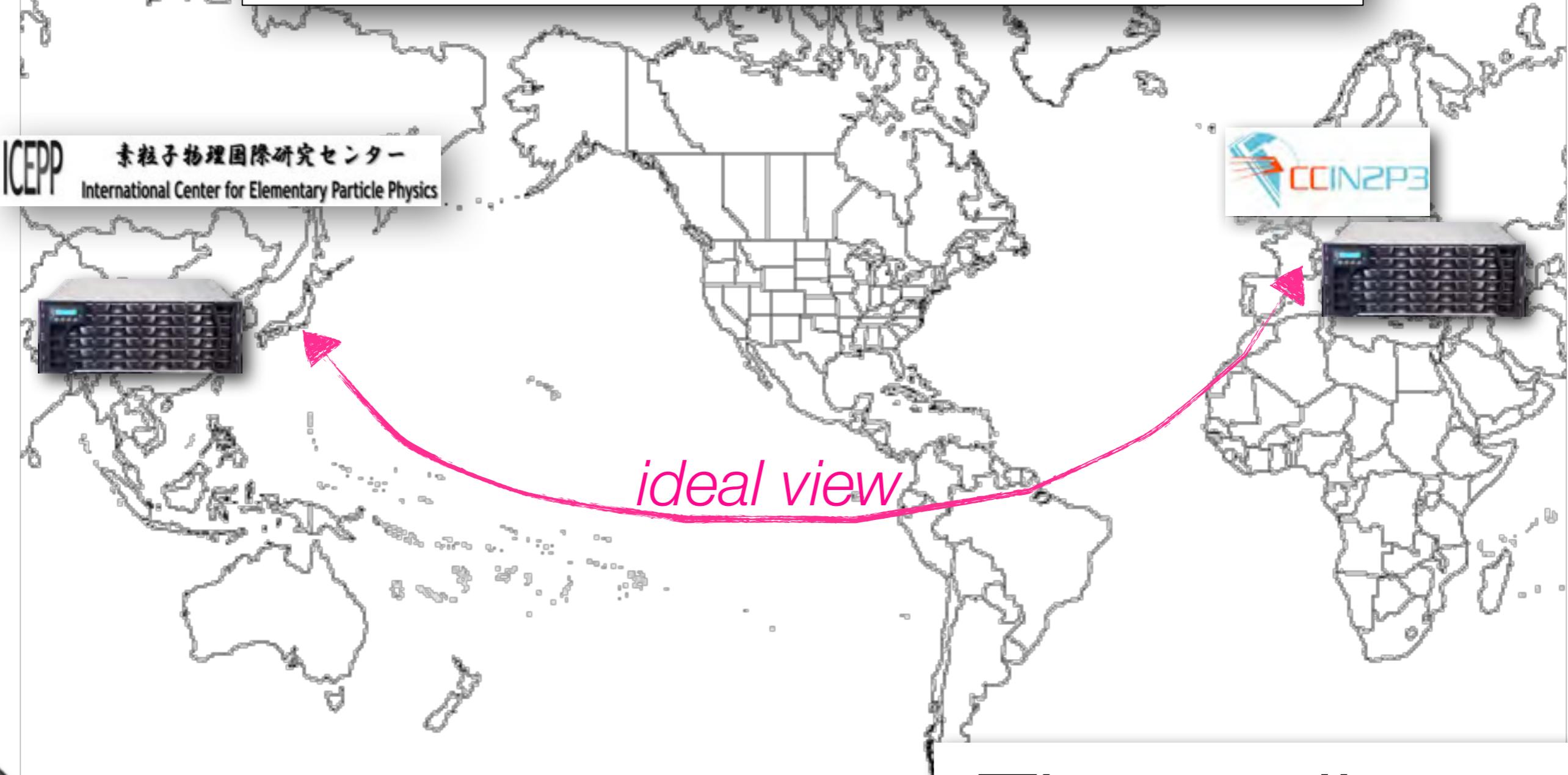
# trans-Siberian route

## GEANT/TEIN3



Tokyo is far from CCIN2P3 : ~300 ms RTT (Round Trip Time)  
Throughput ~ 1 / RTT

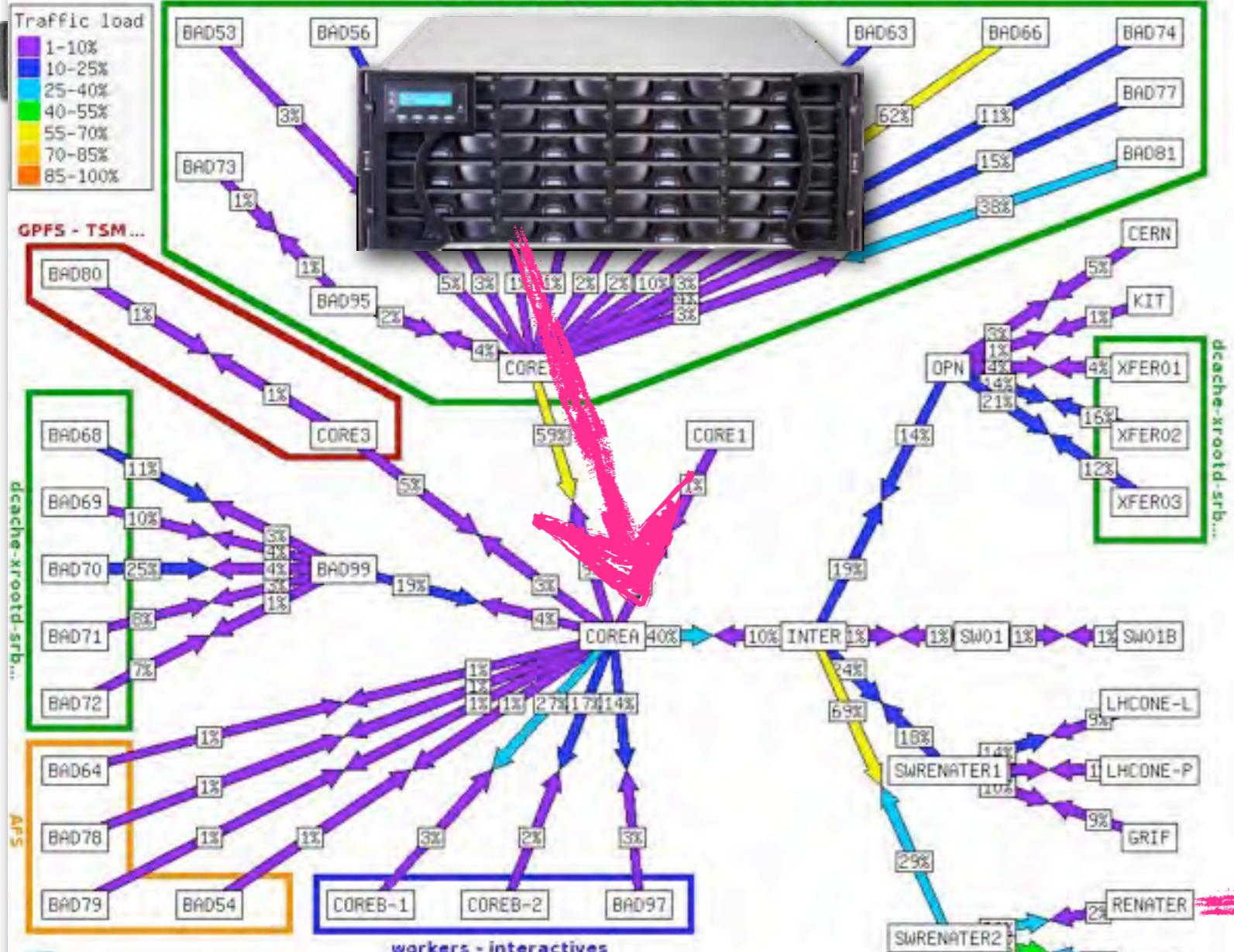
Data are transferred from site to site through a lot of  
networks (multi-hop) and software layers

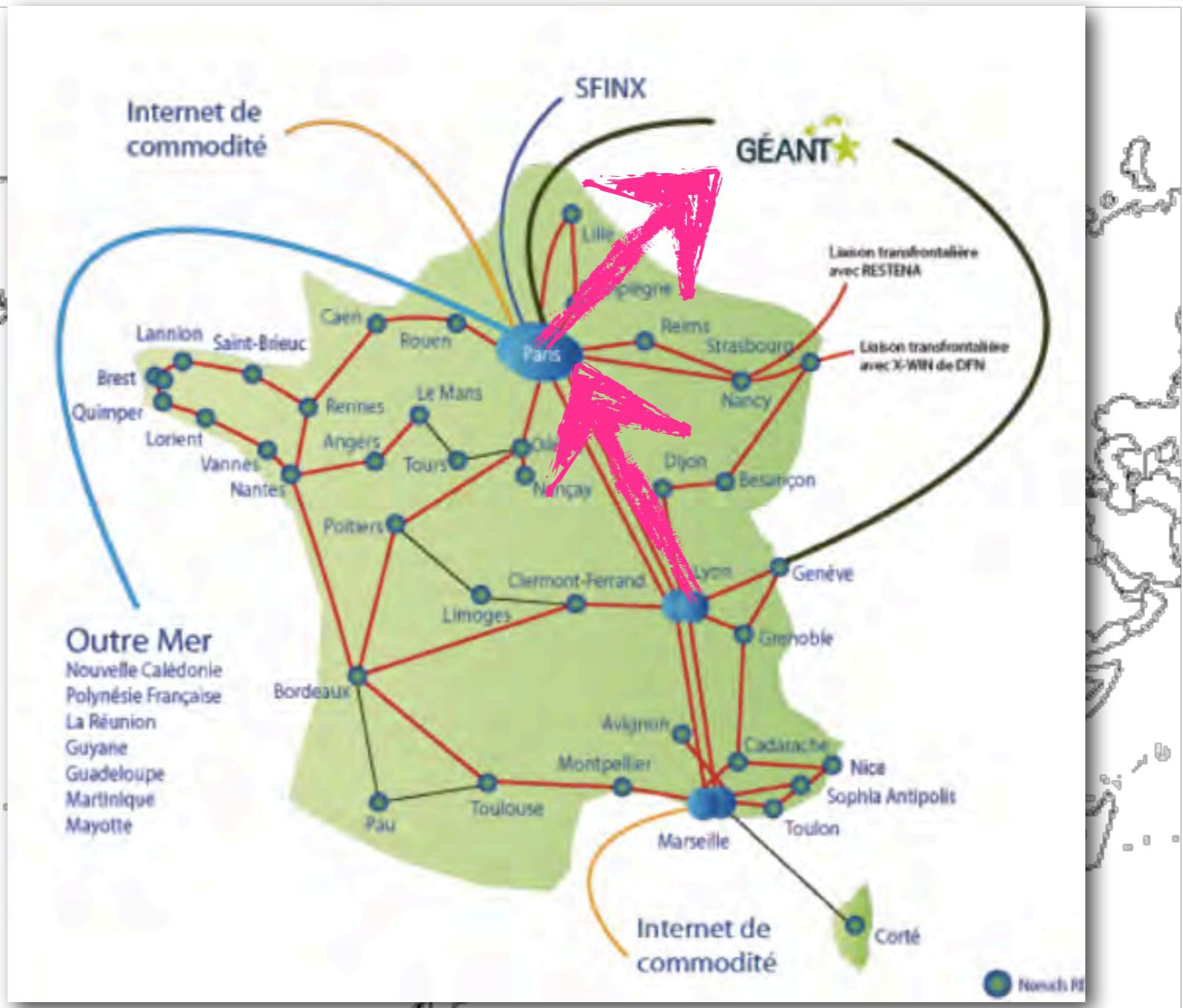


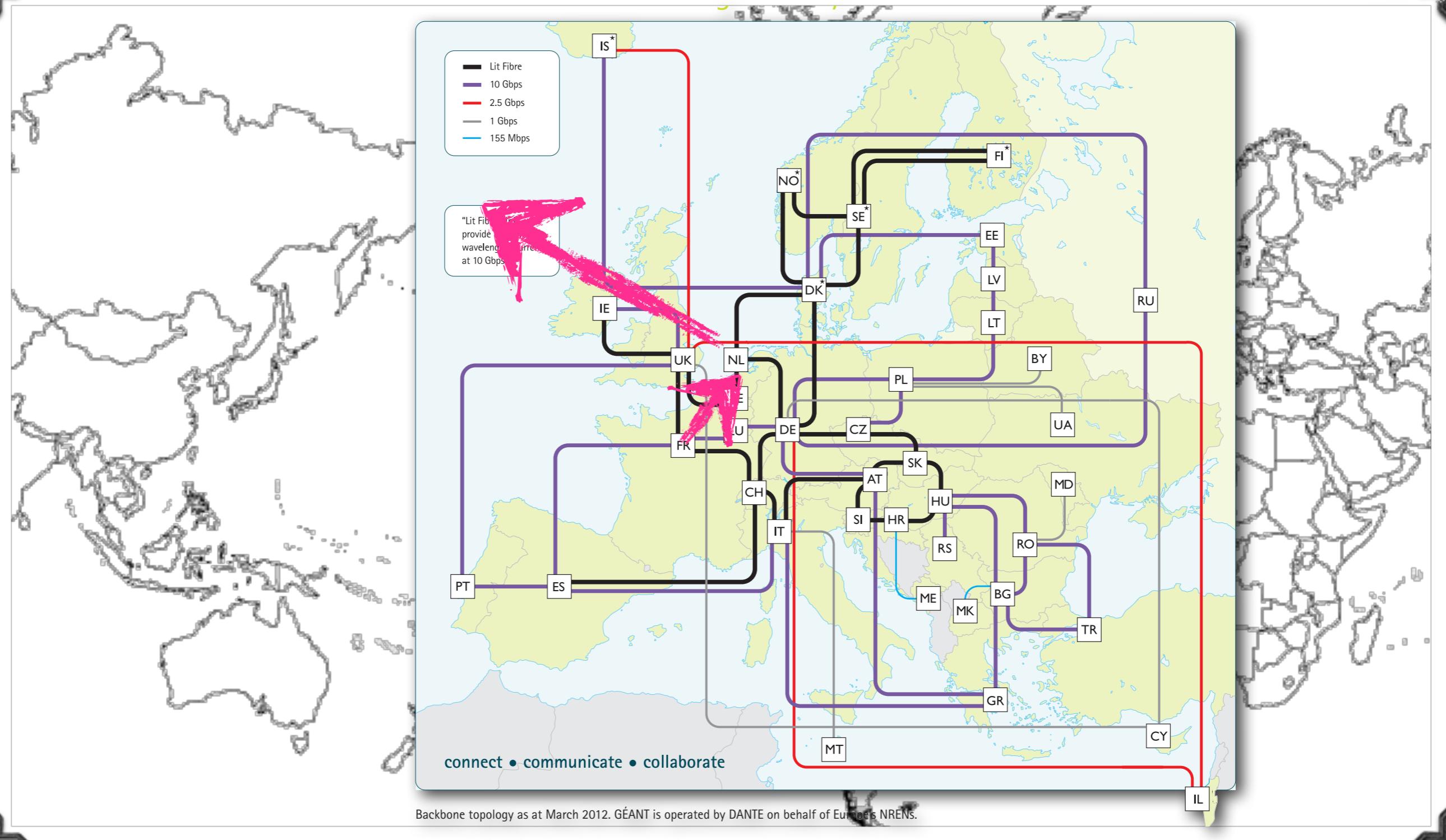
*The reality* ➡

Last update on Thu May 24 16:45:23 2012 UTC

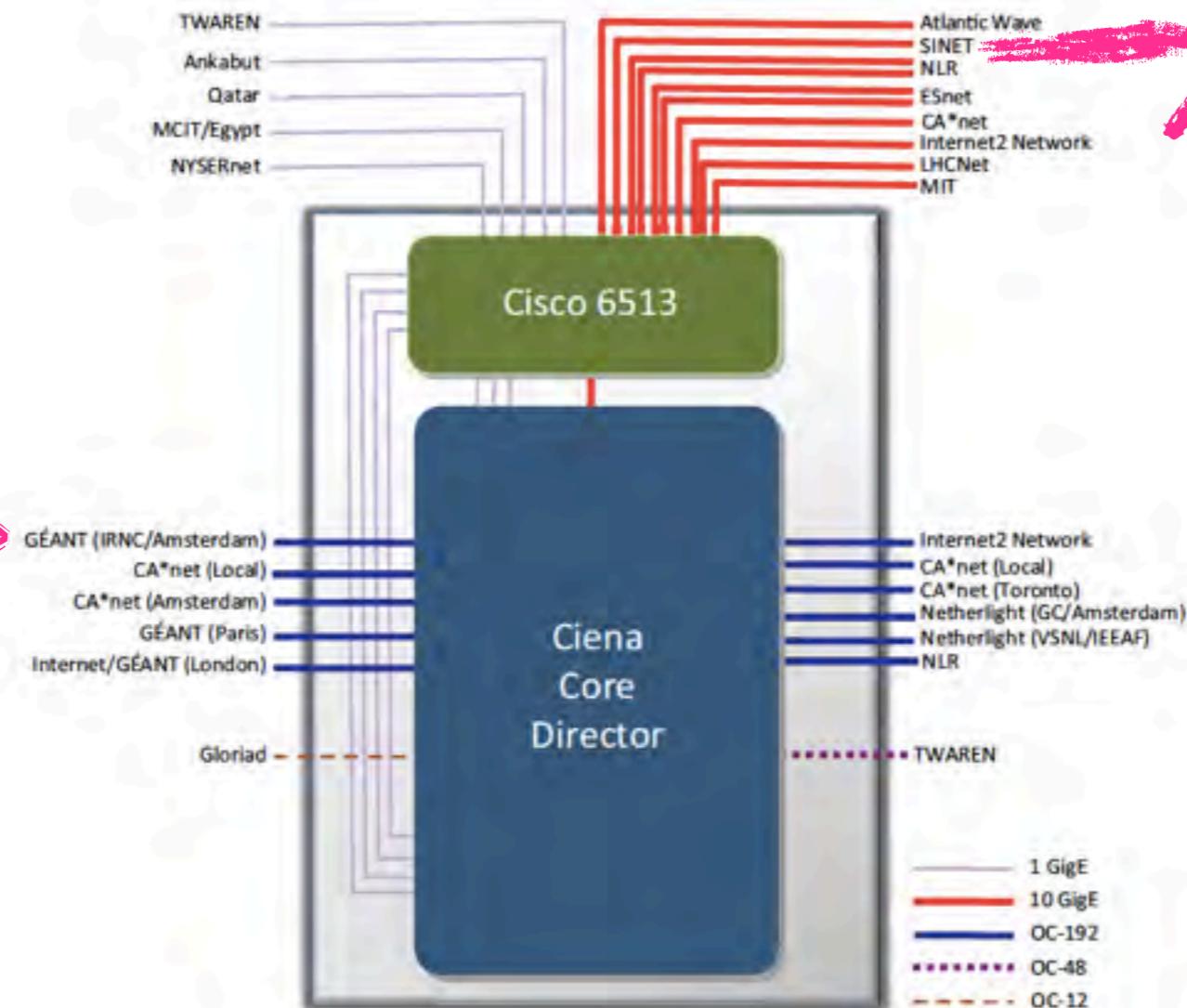
acsis - TSM - HPSS - dcache - xrootd - srb ...

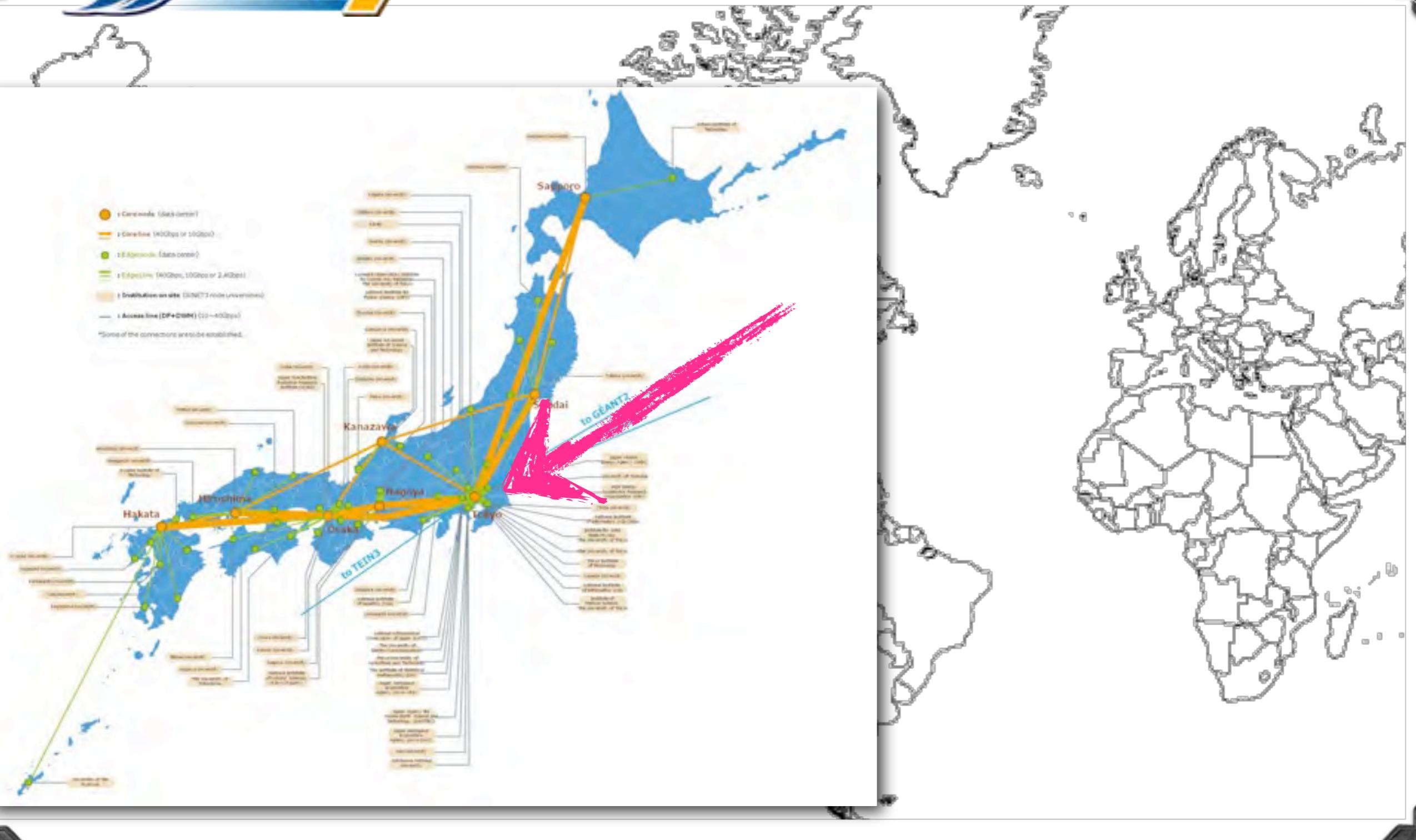


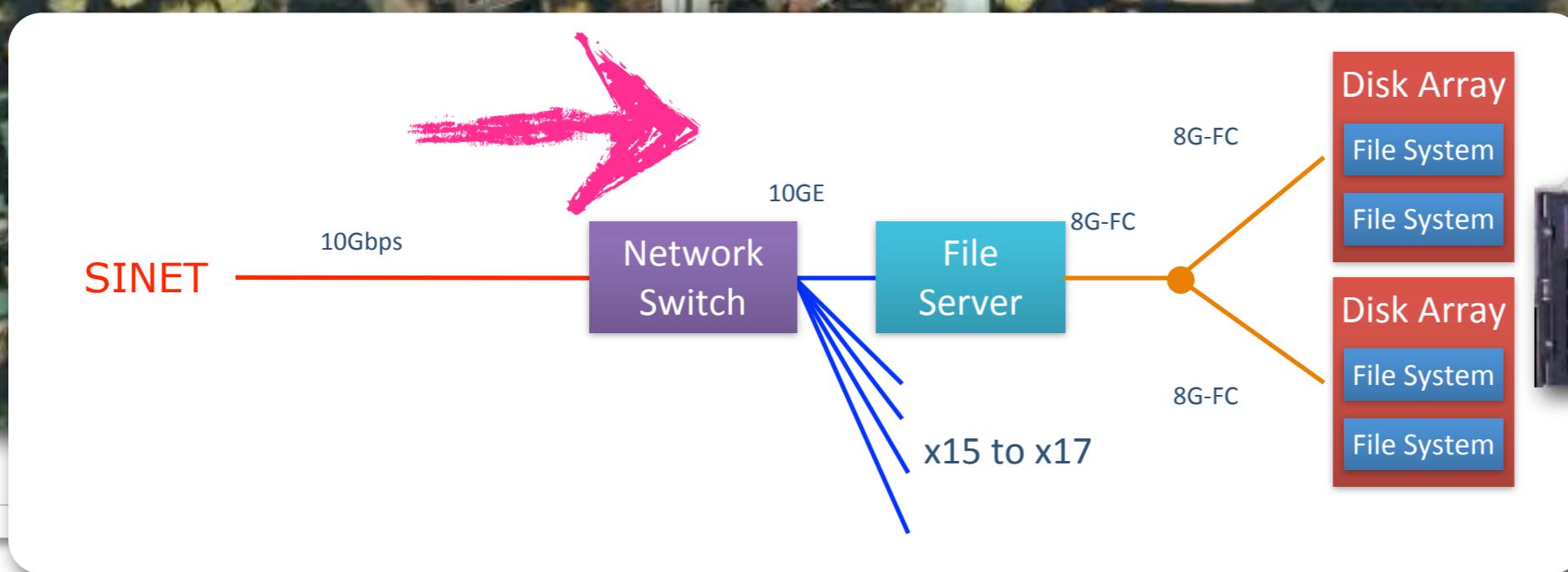
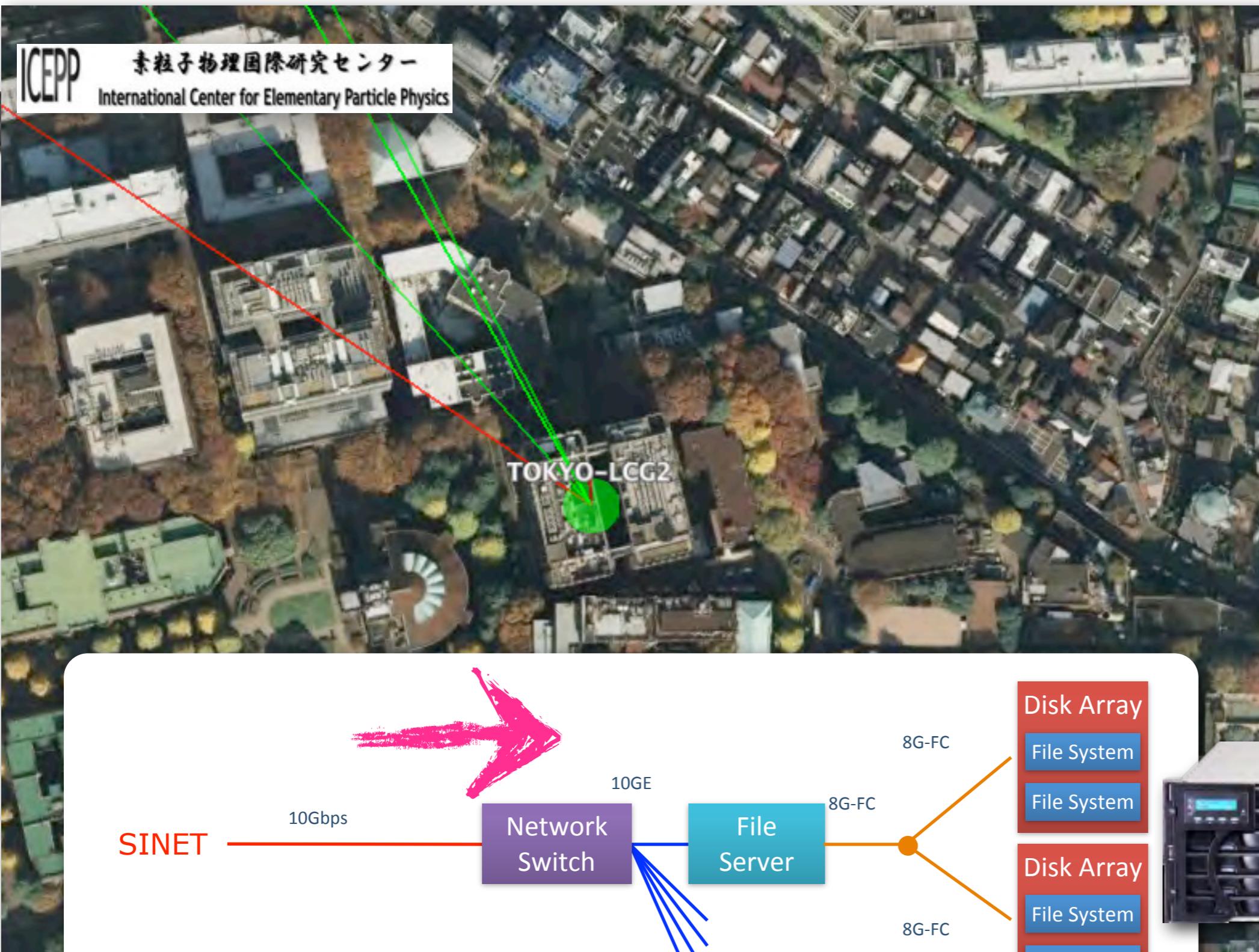




### MAN LAN TOPOLOGY

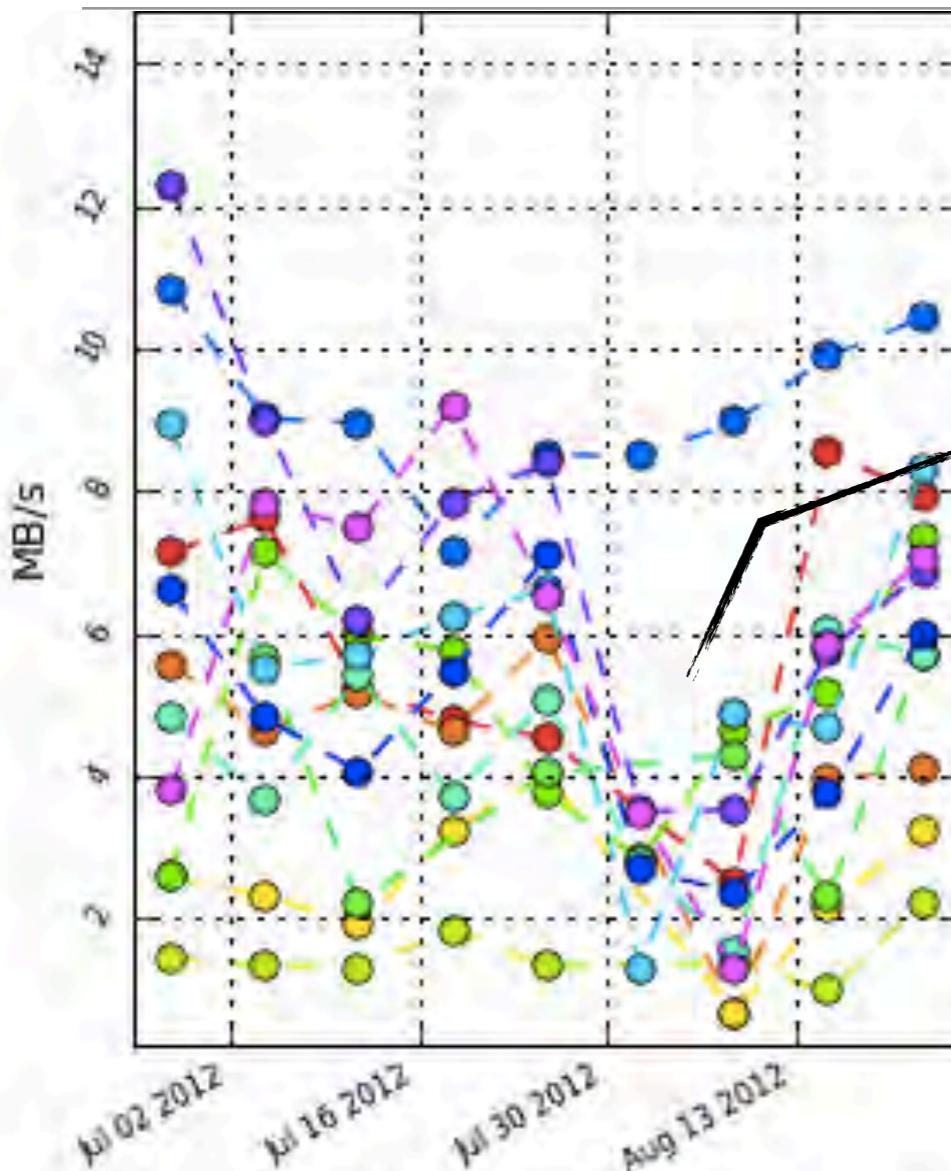




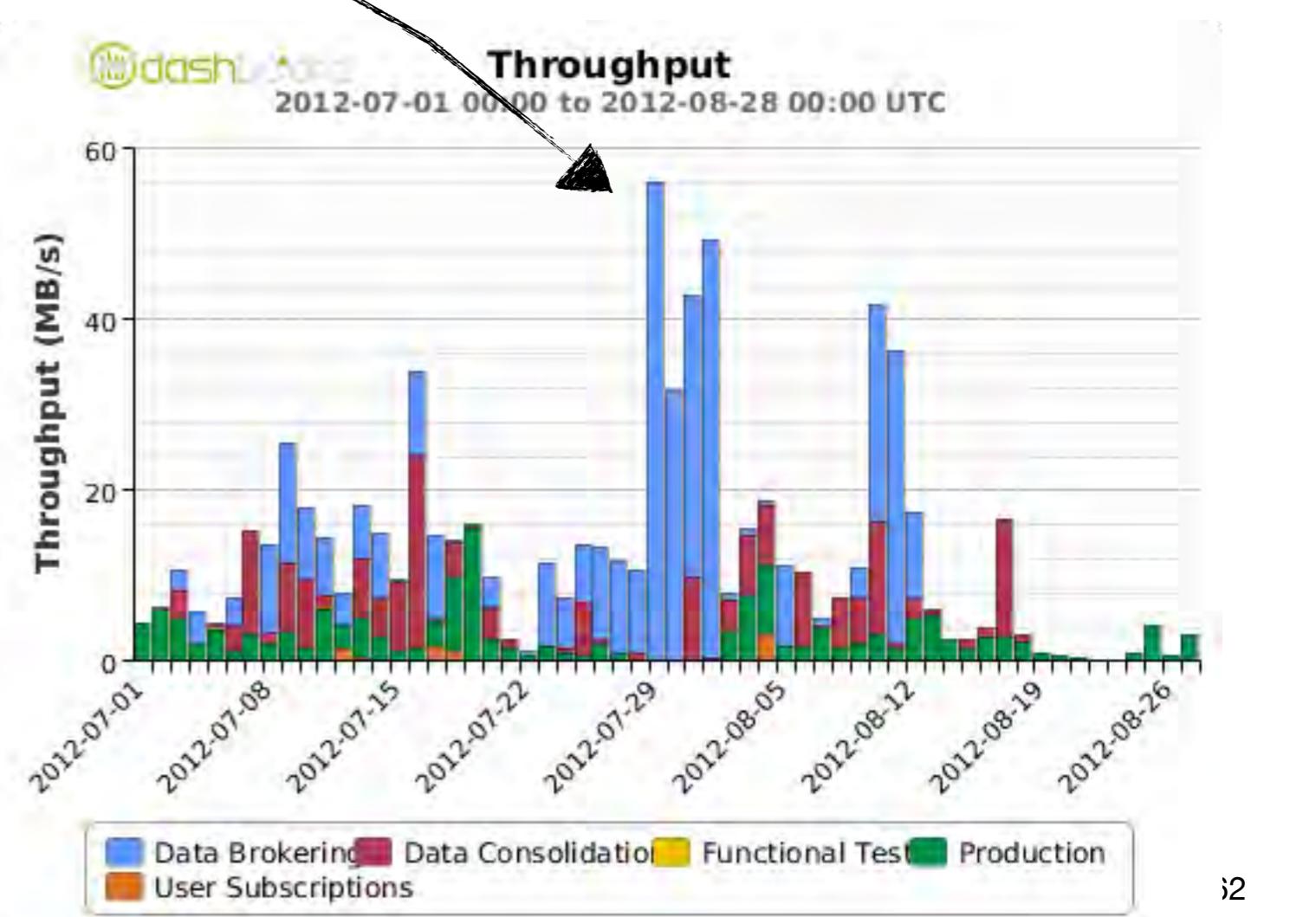


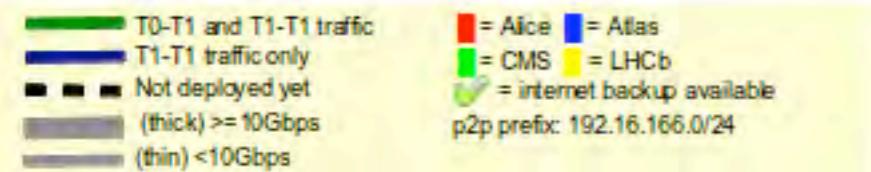
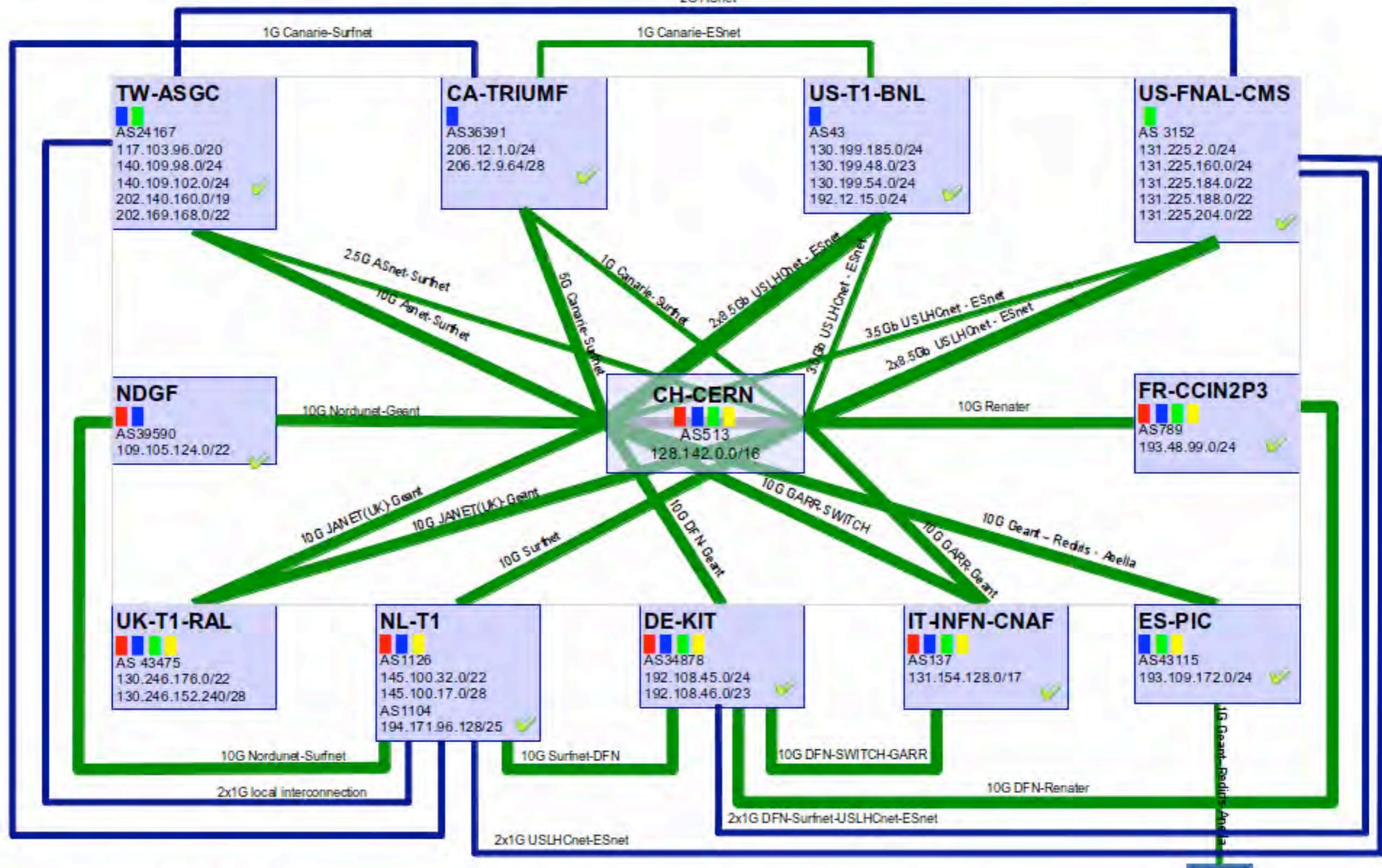
# T1s -> IN2P3-LPSC

*Heavy transfers from IN2P3-CC  
(T1) interference with FTS  
monitoring*



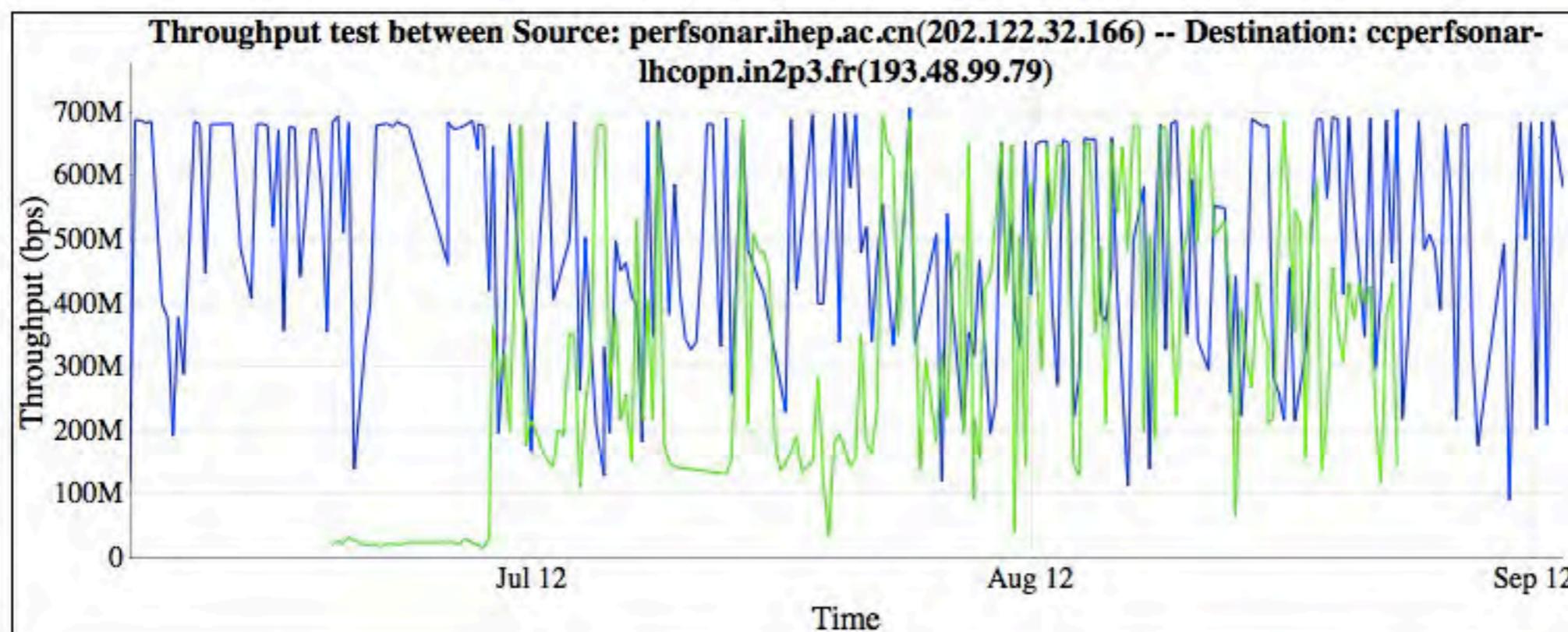
- CERN-PROD\_DATADISK - IN2P3-LPSC (2060 files)
- BNL-OSG2\_DATADISK - IN2P3-LPSC (4210 files)
- TRIUMF-LCG2\_DATADISK - IN2P3-LPSC (2626 files)
- TAIWAN-LCG2\_DATADISK - IN2P3-LPSC (919 files)
- SARA-MATRIX\_DATADISK - IN2P3-LPSC (1989 files)
- NIKHEF-ELPROD\_DATADISK - IN2P3-LPSC (706 files)
- FZK-LCG2\_DATADISK - IN2P3-LPSC (2169 files)
- RAL-LCG2\_DATADISK - IN2P3-LPSC (2255 files)
- IN2P3-CC\_DATADISK - IN2P3-LPSC (21340 files)
- PIC\_DATADISK - IN2P3-LPSC (1525 files)
- INFN-T1\_DATADISK - IN2P3-LPSC (1588 files)
- NDGF-T1\_DATADISK - IN2P3-LPSC (1786 files)





# IN2P3-CC ↔ Beijing as seen by perfSonar

Beijing -> IN2P3-CC    IN2P3-CC -> Beijing



- Link unstable
- Asymmetry

# IN2P3-CC ↔ Tokyo as seen by perfSonar

IN2P3-CC-> Tokyo

Tokyo -> IN2P3-CC

